arXiv:2608.09369v1 [cs.CV] 10 Aug 2026

FeedbackTrack: Visual-Cortex-Inspired Cross-Frame Feedback for Transformer Tracking

Yueyang Cang[1], Xiaoteng Zhang[1], Zhiyuan Ning[1], Yuchen He[1], Li Shi[1]

[1]Tsinghua University
Beijing, China

Abstract

Visual object tracking requires the continuous integration of target information across frames, yet most Transformer trackers still rely on predominantly feed-forward visual feature extraction. Existing temporal mechanisms commonly update templates, prompts, queries, or prediction states, while intermediate representations from previous frames rarely modulate corresponding stages of current-frame processing. We propose **FeedbackTrack**, a visual-cortex-inspired framework that introduces sparse, group-level layer-aligned cross-frame feedback into pretrained Transformer trackers. Previous-frame outputs from selected Transformer groups are detached, cached, and returned to the corresponding groups in the current frame. FeedbackTrack contains two complementary pathways. Query Feedback transforms historical search-token states into low-rank query biases and applies per-sample, per-head, and per-token RMS alignment to control their magnitude. Gate Feedback uses pooled historical context to provide bounded channel-group modulation of the projected attention output. The proposed mechanism retains the original tracking pipeline and requires only a fixed-size one-frame cache. Across SPMTrack and ARTrackV2, FeedbackTrack consistently improves five backbone configurations on LaSOT and GOT-10k. It improves GOT-10k AO by 2.3–4.1 points and LaSOT AUC by 1.1–1.8 points, reaching 83.4 AO and 79.1 AUC with SPMTrack-G, while adding less than 1% parameters in all configurations. Controlled comparisons further show that cross-frame feedback outperforms same-frame modulation by 1.8–3.2 AO points, confirming that the gains primarily arise from recurrent historical information rather than additional modulation capacity. Analysis of the learned feedback scales reveals a non-uniform depth-dependent organization, demonstrating that recurrent cross-frame feedback provides an effective and scalable mechanism for Transformer tracking.

Introduction

Visual object tracking aims to continuously localize an arbitrary target throughout a video given only its initial state. A reliable tracker must preserve target identity while adapting to appearance variation, occlusion, fast motion, background clutter, and similar distractors. Recent advances have shifted tracking architectures from convolutional Siamese matching toward Transformer-based target–search interaction (Li et al. 2019; Chen et al. 2021; Yan et al. 2021; Cui et al. 2022). In particular, one-stream and sequence-based trackers jointly model template and search information within a shared backbone, enabling strong feature interaction and accurate target localization (Ye et al. 2022; Chen et al. 2023; Wei et al. 2023; Cai, Liu, and Wang 2025).

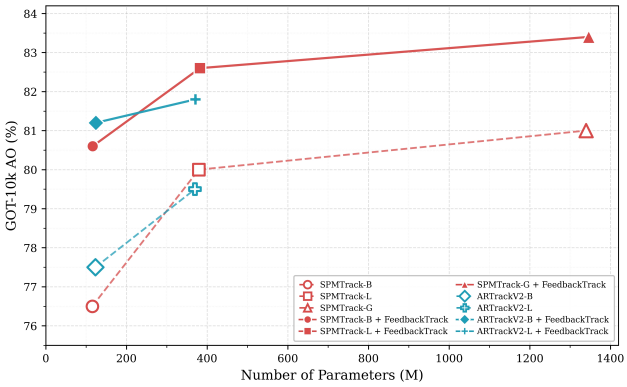


Figure 1: Performance versus model parameter count on GOT-10k for the SPMTrack and ARTrackV2 series. Solid and dashed lines denote the FeedbackTrack-enhanced and original models, respectively. FeedbackTrack consistently improves tracking performance across different backbone scales with only marginal parameter overhead.

Despite their strong performance, existing Transformer trackers still rely predominantly on feed-forward visual feature extraction within each frame. Temporal information has been incorporated through updated templates, historical prompts, autoregressive queries, candidate associations, and additional temporal tokens (Wang et al. 2021; Mayer et al. 2021; Cai, Liu, and Wang 2024; Zheng et al. 2024; Xie et al. 2024; Bai et al. 2024). These mechanisms effectively preserve target history, but temporal states are commonly introduced at the input, prompt, query, or prediction level and then processed by another largely feed-forward visual encoder. Intermediate representations formed at a particular backbone stage in the previous frame are rarely returned to the corresponding processing stage in the current frame. Consequently, the visual feature hierarchy itself lacks an explicit recurrent pathway through which previously formed representations can directly modulate ongoing feature extraction.

Biological vision suggests a different computational orga-

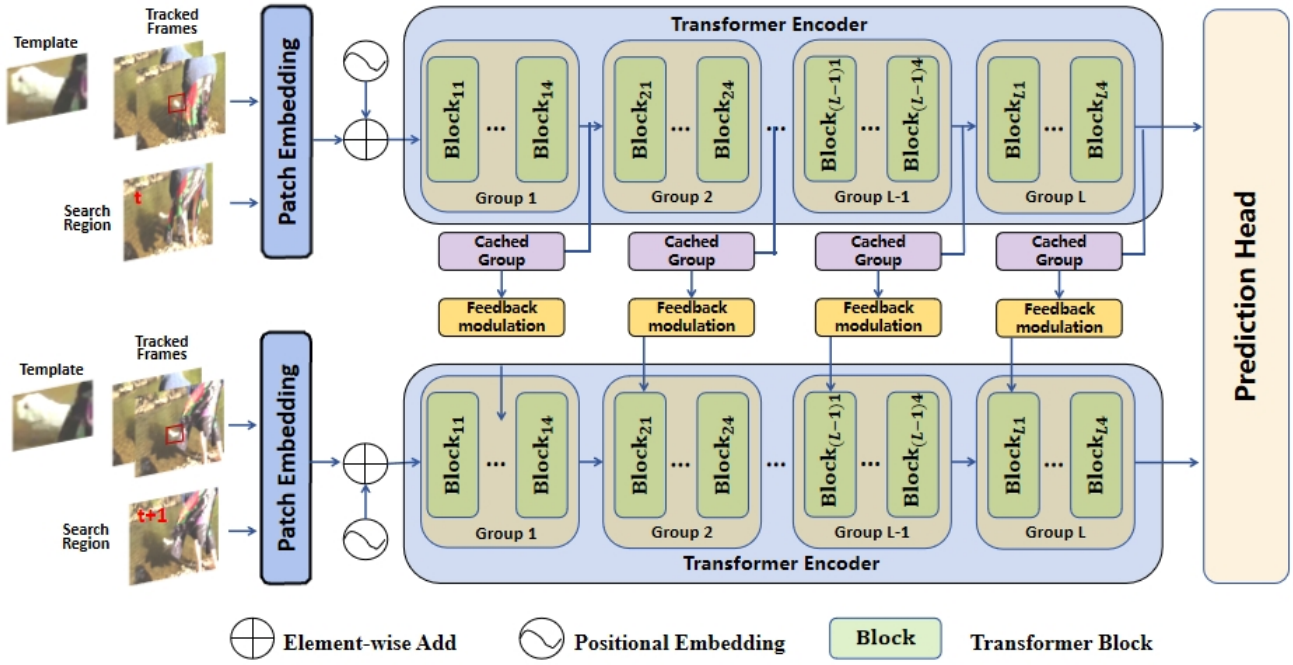


Figure 2: Overall architecture of FeedbackTrack. Previous-frame outputs of selected Transformer groups are detached, cached, and returned to the corresponding groups in the current frame. The original prediction head remains unchanged.

nization. The mammalian visual system combines ascending feed-forward pathways with extensive recurrent and feedback connections, allowing previously formed representations to influence subsequent sensory processing (Lamme and Roelfsema 2000; Markov et al. 2014). Recent studies further indicate that visual feedback is distributed across processing stages, pathway-specific, and modulatory rather than a simple reversal of feed-forward activity (Federer et al. 2021; Siu et al. 2021; Semedo et al. 2022; Fişek et al. 2023; Shen et al. 2022). These findings motivate us to investigate whether recurrent reuse of intermediate visual states can improve Transformer tracking. We do not seek to reproduce specific cortical structures or neural dynamics, but instead abstract the general principle that previously formed representations can provide stage-corresponding modulation of current visual processing.

To this end, we propose **FeedbackTrack**, a visual-cortex-inspired framework that introduces sparse, group-level layer-aligned cross-frame feedback into pretrained Transformer trackers. The visual encoder is partitioned into groups of consecutive Transformer blocks. For each selected group, the previous-frame group output is detached, cached, and returned to the first block of the corresponding group in the current frame. This design converts the originally stateless feature hierarchy into a lightweight recurrent system while retaining the original feed-forward pathway, tracking interface, and prediction modules. Only one-frame states are retained, so the cache size remains constant with respect to video length.

FeedbackTrack implements recurrent modulation through two complementary pathways. **Query Feedback** transforms previous search-token states through a low-rank projection to generate token-wise biases for current search queries. A per-sample, per-head, and per-token RMS alignment operation controls the bias magnitude relative to the pretrained queries. **Gate Feedback** pools the complete previous group state and generates bounded, context-dependent channel-group modulation of the projected attention output. Query Feedback therefore influences where current attention retrieves evidence, whereas Gate Feedback regulates the resulting attention response.

We instantiate FeedbackTrack on two representative tracking frameworks, SPMTrack and ARTrackV2. On SPMTrack, FeedbackTrack improves GOT-10k AO from 76.5 to 80.6 for ViT-B, from 80.0 to 82.6 for ViT-L, and from 81.0 to 83.4 for ViT-G. The corresponding LaSOT AUC scores increase from 74.9 to 76.4, from 76.8 to 77.9, and from 77.4 to 79.1. FeedbackTrack also improves ARTrackV2-B and ARTrackV2-L to 81.2 and 81.8 AO on GOT-10k, with corresponding LaSOT AUC scores of 74.8 and 75.4. Across all five configurations, the additional parameter cost remains below 1%. Moreover, under the same modulation architecture and parameter count, cross-frame feedback outperforms its same-frame counterpart by 3.2, 2.3, and 1.8 AO points for ViT-B, ViT-L, and ViT-G, respectively, demonstrating that the improvements primarily arise from recurrent historical information rather than additional modulation capacity alone.

Although all Query Feedback scales are initialized identically, the trained models develop a non-uniform depth-

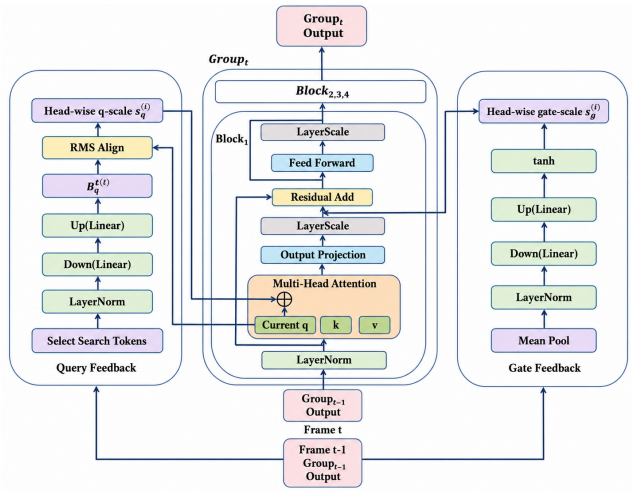


Figure 3: Feedback module. Query Feedback generates a norm-controlled bias for current search-token queries, while Gate Feedback modulates the projected attention output using pooled previous-frame context.

dependent organization, with relatively weak feedback in shallow groups and stronger modulation in intermediate and deep groups. This emergent pattern is qualitatively consistent with the hierarchical and pathway-dependent organization of biological visual feedback, while representing a computational correspondence rather than an anatomical equivalence.

Our main contributions are summarized as follows:

- We introduce FeedbackTrack, a general framework that equips pretrained Transformer trackers with sparse, group-level layer-aligned cross-frame feedback, enabling intermediate previous-frame states to recurrently modulate corresponding stages of current-frame visual processing.
- We develop two complementary lightweight pathways. Query Feedback performs low-rank, RMS-aligned modulation of current search queries, while Gate Feedback applies bounded, context-dependent channel-group rescaling to the projected attention output.
- We demonstrate consistent improvements across two tracking frameworks and five backbone configurations on LaSOT and GOT-10k, while introducing less than 1% additional parameters. Controlled same-frame comparisons further verify the importance of recurrent historical information.
- We analyze the learned feedback strengths and identify an emergent depth-dependent organization, showing that cross-frame feedback is allocated non-uniformly across visual representation stages.

Related Work

Transformer-Based Visual Tracking

Visual object tracking has progressively shifted from convolutional Siamese matching to Transformer-based architectures with explicit target–search interaction. Representative methods include TransT, STARK, OSTrack, SeqTrack, ARTrack, and ARTrackV2 (Chen et al. 2021; Yan et al. 2021; Ye et al. 2022; Chen et al. 2023; Wei et al. 2023; Bai et al. 2024), while more recent approaches further explore adaptive temporal queries and parameter-efficient tracking (Xie et al. 2024; Cai, Liu, and Wang 2025). Although these trackers may use historical templates, prompts, or autoregressive predictions, intermediate visual representations are generally not recurrently returned to corresponding backbone stages in the following frame. FeedbackTrack complements these architectures by introducing sparse, group-level layer-aligned cross-frame feedback inside the visual encoder.

Temporal Modeling in Visual Tracking

Existing trackers exploit temporal context through historical templates, candidate associations, memory representations, prompts, autoregressive queries, and temporal tokens (Wang et al. 2021; Mayer et al. 2021; Cai, Liu, and Wang 2024; Zheng et al. 2024; Xie et al. 2024; Bai et al. 2024). These approaches typically introduce temporal information at the input, prompt, memory, or prediction level, after which the visual encoder remains predominantly feed-forward. In contrast, FeedbackTrack treats intermediate Transformer group outputs as recurrent visual states and returns each previous-frame group output to the corresponding current-frame group. This design is complementary to existing temporal modeling strategies, including the prompt-based autoregressive pipeline of ARTrackV2.

Recurrent Feedback in Biological Vision

Biological vision combines feed-forward and recurrent processing, with distributed and pathway-specific feedback modulating ongoing sensory representations across the visual hierarchy (Lamme and Roelfsema 2000; Markov et al. 2014; Federer et al. 2021; Semedo et al. 2022; Fişek et al. 2023; Shen et al. 2022). FeedbackTrack does not reproduce specific cortical structures or neural pathways, but abstracts several general computational principles, including recurrent reuse of previous representations, correspondence between feedback states and processing stages, and residual modulation of current visual computation. These principles are implemented through group-level layer-aligned cross-frame connections together with Query Feedback and Gate Feedback.

Method

Overview

Biological visual processing combines feed-forward computation with recurrent feedback that modulates ongoing sensory representations (Federer et al. 2021; Siu et al. 2021; Semedo et al. 2022; Fişek et al. 2023; Shen et al. 2022). Inspired by this computational principle, FeedbackTrack introduces cross-frame recurrent modulation into the visual Transformer encoder, without attempting to reproduce specific cortical structures.

FeedbackTrack is formulated as a generic feedback mechanism for Transformer-based trackers and is instantiated on two representative frameworks, SPMTrack (Cai, Liu, and

Wang 2025) and ARTrackV2 (Bai et al. 2024). For SPMTrack, the feedback modules are inserted into its DINOv2-based Transformer backbone (Oquab et al. 2024; Dosovitskiy et al. 2021), while the patch embedding, tracking embeddings, TMoE adapters, and prediction head remain unchanged. For ARTrackV2, only the visual encoder is modified, while the original trajectory prompts, appearance prompts, autoregressive prediction modules, and training objective are retained. As illustrated in Fig. 2, FeedbackTrack augments the original visual feature extractor with sparse cross-frame connections while preserving the task-specific tracking interface and prediction pipeline.

Layer-Aligned Cross-Frame Feedback

We partition the Transformer backbone into groups of G consecutive blocks. Let $\mathbf{X}_t^{(i)}$ and $\mathbf{U}_t^{(i)}$ denote the input and output of group i at frame t, respectively. The output of the corresponding group in the previous frame is detached and stored as a one-frame cache:

$$\mathbf{C}_{t-1}^{(i)} = \mathrm{StopGrad}\left(\mathbf{U}_{t-1}^{(i)}\right). \tag{1}$$

At frame t, the cached representation is provided to the first Transformer block of the same group:

$$\mathbf{U}_t^{(i)} = \mathcal{G}_{\mathrm{fb}}^{(i)}\left(\mathbf{X}_t^{(i)}, \mathbf{C}_{t-1}^{(i)}\right), \tag{2}$$

where $\mathcal{G}_{\mathrm{fb}}^{(i)}$ denotes group i equipped with the proposed feedback module.

Because the output of group i at frame $t-1$ is returned to the first block of the same group i at frame t, the temporal connection preserves the hierarchical stage at which the representation was generated. We therefore refer to this design as *group-level layer-aligned cross-frame feedback*. Only the first block of each group contains the feedback module, whereas the remaining blocks retain their original computation.

We use a group size of $G = 4$. For example, the output produced after blocks 0–3 at frame $t-1$ is returned to block 0 at frame t, while the output produced after blocks 4–7 is returned to block 4. Accordingly, feedback modules are inserted at blocks $\{0, 4, 8\}$ for ViT-B, $\{0, 4, 8, 12, 16, 20\}$ for ViT-L, and $\{0, 4, 8, 12, 16, 20, 24, 28, 32, 36\}$ for ViT-G. The same grouping strategy is applied to the visual encoders of ARTrackV2-B and ARTrackV2-L.

Feedback Module

The feedback module contains two complementary pathways, as illustrated in Fig. 3. Query Feedback preserves token-level spatial information and recurrently modulates the current search-token queries. Gate Feedback summarizes the complete previous group state and applies context-dependent channel-group modulation to the projected attention output. The two pathways therefore inject historical information at different locations in the attention computation.

Norm-Controlled Query Feedback Let

$$\mathbf{M}_{t-1}^{q,(i)} \in \mathbb{R}^{B\times N_s\times D} \tag{3}$$

denote the search-token subset selected from the cached representation $\mathbf{C}_{t-1}^{(i)}$, where B is the batch size, N_s is the number of search tokens, and D is the feature dimension. We first transform the historical search-token states using LayerNorm and a low-rank projection:

$$\mathbf{B}_t^{q,(i)} = \mathrm{Up}^{(i)}\left(\mathrm{Down}^{(i)}\left(\mathrm{LN}\left(\mathbf{M}_{t-1}^{q,(i)}\right)\right)\right), \tag{4}$$

where the down- and up-projections map the feature dimension as $D \to r \to D$, with a bottleneck dimension of $r = 16$. The projected historical bias is then reshaped into the multi-head form

$$\mathbf{B}_t^{q,(i)} \in \mathbb{R}^{B\times H\times N_s\times d_h}, \tag{5}$$

where H is the number of attention heads and $D = Hd_h$.

Directly adding the projected historical bias to the pretrained queries may result in a substantial magnitude mismatch. We therefore align their root-mean-square magnitudes independently for each sample, attention head, and search token:

$$\widehat{\mathbf{B}}_{t,bhs:}^{q,(i)} = \mathbf{B}_{t,bhs:}^{q,(i)} \frac{\rho\ \mathrm{RMS}\left(\mathbf{Q}_{t,bhs:}^{(i)}\right)}{\mathrm{RMS}\left(\mathbf{B}_{t,bhs:}^{q,(i)}\right)}, \tag{6}$$

where b, h, and s index the sample, attention head, and search token, respectively. The RMS operator is defined as

$$\mathrm{RMS}(\mathbf{x}) = \sqrt{\frac{1}{d_h}\sum_{d=1}^{d_h} x_d^2 + \epsilon}. \tag{7}$$

We set $\rho = 0.05$ and $\epsilon = 10^{-4}$. The aligned historical bias is added to the current search-token queries using a learnable head-wise scale vector $\mathbf{s}_q^{(i)} \in \mathbb{R}^H$:

$$\widetilde{\mathbf{Q}}_{t,bhs:}^{(i)} = \mathbf{Q}_{t,bhs:}^{(i)} + s_{q,h}^{(i)}\widehat{\mathbf{B}}_{t,bhs:}^{q,(i)}. \tag{8}$$

All q-scales are initialized to 0.01, allowing the modified tracker to start close to the pretrained model. Only the current search-token queries are modified, while all non-search queries, keys, and values remain unchanged.

Gate Feedback Query Feedback is complemented by an output-side modulation pathway. We first summarize the complete previous group state by averaging over all cached tokens:

$$\mathbf{p}_{t-1}^{(i)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{C}_{t-1,n}^{(i)}, \tag{9}$$

where $\mathbf{p}_{t-1}^{(i)} \in \mathbb{R}^{B\times D}$ and N denotes the total number of cached tokens.

A lightweight MLP maps the pooled context from D to a hidden dimension of 384 and then to H channel groups. A

tanh-bounded contextual signal is combined with a learnable channel-group scale vector $\mathbf{s}_g^{(i)} \in \mathbb{R}^H$:

$$\mathbf{g}_t^{(i)} = \tanh\left(\mathrm{MLP}_g^{(i)}\left(\mathrm{LN}\left(\mathbf{p}_{t-1}^{(i)}\right)\right)\right) \odot \mathbf{s}_g^{(i)}, \tag{10}$$

where $\mathbf{g}_t^{(i)} \in \mathbb{R}^{B\times H}$ contains one modulation coefficient for each channel group. The learnable gate scales are initialized to 0.01.

Let $\mathbf{A}_t^{(i)}$ denote the attention output after head concatenation, output projection, projection dropout, and the original LayerScale. We reshape its feature dimension into H channel groups, each containing d_h channels, and apply

$$\widetilde{\mathbf{A}}_{t,bnh:}^{(i)} = \left(1 + g_{t,bh}^{(i)}\right) \mathbf{A}_{t,bnh:}^{(i)}. \tag{11}$$

The same gate is broadcast across all tokens and applied before the original residual connection. Because the output projection has already mixed information from the original attention heads, this operation should be interpreted as channel-group rescaling of the projected attention output rather than direct modulation of individual attention heads.

Training and Online Inference

During training, consecutive search frames are processed sequentially, and detached previous-frame group outputs serve as feedback caches for the current frame, avoiding cross-frame backpropagation through time. The original feed-forward and residual pathways are retained, while the small initial query and gate scales keep the model close to the pretrained base tracker.

During inference, each selected Transformer group maintains a one-frame cache that is updated after every frame. When no valid cache is available, the original base-tracker computation is used. Because only the immediately preceding group states are stored, the cache memory remains constant with respect to video length.

Experiments

Datasets and Evaluation Metrics

We evaluate FeedbackTrack on two widely used visual tracking benchmarks, LaSOT (Fan et al. 2019) and GOT-10k (Huang, Zhao, and Huang 2019). For LaSOT, we report the area under the success curve (AUC), normalized precision (P_{Norm}), and precision (P). For GOT-10k, we report average overlap (AO) and success rates at overlap thresholds of 0.5 and 0.75, denoted as $\mathrm{SR}_{0.5}$ and $\mathrm{SR}_{0.75}$, respectively.

Implementation Details

We evaluate FeedbackTrack on two representative Transformer tracking frameworks, SPMTrack (Cai, Liu, and Wang 2025) and ARTrackV2 (Bai et al. 2024). For SPMTrack, FeedbackTrack is implemented in TrackIt and evaluated with ViT-B, ViT-L, and ViT-G backbones. For ARTrackV2-B and ARTrackV2-L, the feedback modules are inserted only into the visual encoder, while the original prompts, autoregressive prediction modules, and training objectives remain unchanged.

The SPMTrack-based models are trained with AdamW on four NVIDIA A100-SXM4-80GB GPUs using a total batch size of 32. The base learning rate is 1×10^{-5} with a weight decay of 0.1. We use a per-iteration cosine schedule with a 10-epoch warm-up from 1×10^{-7} and a minimum learning rate of 1×10^{-6}. Most backbone parameters are frozen, while the last four Transformer blocks are fine-tuned at 1×10^{-6}. The TMoE adapters, tracking embeddings, and prediction head use the base learning rate, whereas Query Feedback and Gate Feedback are optimized at 5×10^{-5}. For ARTrackV2, we follow its original training configuration.

Inference throughput is measured on a single NVIDIA A100 GPU with batch size 1. FPS is averaged over all evaluated GOT-10k test frames, excluding image loading and preprocessing.

Main Results

Table 1 compares FeedbackTrack with representative state-of-the-art trackers on LaSOT and GOT-10k. When applied to SPMTrack, FeedbackTrack consistently improves all six metrics across the B, L, and G variants. On GOT-10k, the AO scores increase by 4.1, 2.6, and 2.4 points, respectively, while the corresponding LaSOT AUC gains are 1.5, 1.1, and 1.7 points.

FeedbackTrack also generalizes well to ARTrackV2. It improves ARTrackV2-B and ARTrackV2-L to 81.2 and 81.8 AO on GOT-10k, with LaSOT AUC scores of 74.8 and 75.4. The corresponding $\mathrm{SR}_{0.75}$ gains are 5.8 and 2.5 points, showing that intermediate visual-state feedback complements the autoregressive temporal modeling of ARTrackV2.

Ablation Study

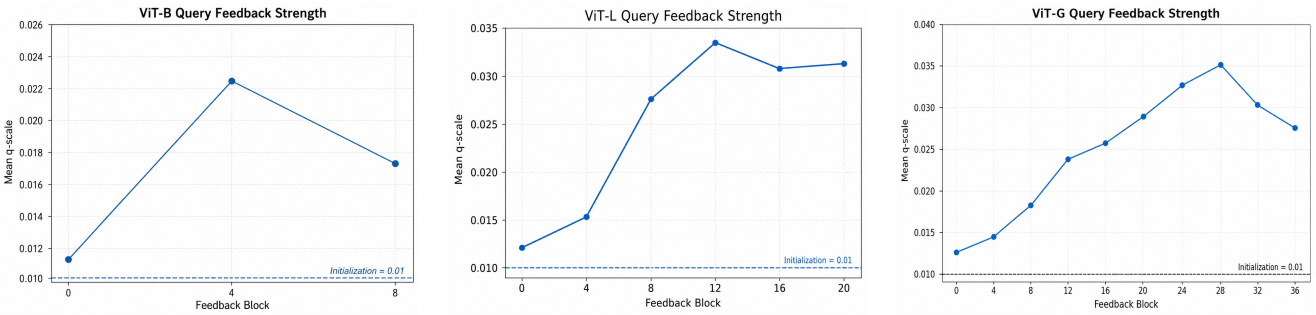
Effect of feedback components. Table 2 shows that both Query Feedback and Gate Feedback are independently effective. Query Feedback contributes the majority of the gain, improving AO by 3.2, 2.0, and 2.0 points for ViT-B, ViT-L, and ViT-G, while Gate Feedback yields gains of 1.8, 0.6, and 0.8 points, respectively. Combining both pathways achieves the best results on all metrics, indicating that token-level query modulation and global context-based output modulation are complementary.

Cross-frame feedback versus same-frame modulation. To isolate the effect of recurrent historical information, we construct a same-frame control that replaces the detached previous-frame group output with the detached current group input, while keeping the same modules, insertion positions, and parameter count. As shown in Table 3, cross-frame feedback outperforms this control by 3.2, 2.3, and 1.8 AO points for ViT-B, ViT-L, and ViT-G, respectively, confirming that the main gains arise from recurrent historical information rather than additional modulation capacity.

Effect of RMS alignment. Table 4 evaluates RMS alignment in Query Feedback. Aligning the historical bias with the magnitude of current search queries improves all metrics, increasing AO by 2.4, 0.5, and 1.3 points for ViT-B, ViT-L, and ViT-G, respectively. This confirms the importance of controlling feedback magnitude when modulating pretrained query representations.

Table 1: Comparison with state-of-the-art trackers on LaSOT and GOT-10k. Each indented FeedbackTrack row is built upon the immediately preceding baseline. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Source	LaSOT			GOT-10k		
		AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$
AiATrack	ECCV'22	69.0	79.4	73.8	69.6	80.0	63.2
OSTrack_{384}	ECCV'22	71.1	81.1	77.6	73.7	83.2	70.8
ARTrack_{384}	CVPR'23	72.6	81.7	79.1	75.5	84.3	74.3
GRM	CVPR'23	69.9	79.3	75.8	73.4	82.9	70.4
SeqTrack-B_{384}	CVPR'23	71.5	81.1	77.8	74.5	84.3	71.4
F-BDMTrack_{384}	ICCV'23	72.0	81.5	77.7	75.4	84.3	72.9
ROMTrack_{384}	ICCV'23	71.4	81.4	78.2	74.2	84.3	72.4
ODTrack-B	AAAI'24	73.2	83.2	80.6	77.0	87.9	75.1
AQATrack_{384}	CVPR'24	72.7	82.9	80.2	76.0	85.2	74.9
ARTrackV2-B_{384}	CVPR'24	73.0	82.0	79.6	77.5	86.0	75.5
+ FeedbackTrack	Ours	74.8	82.7	80.2	81.2	89.2	81.3
ARTrackV2-L_{384}	CVPR'24	73.6	82.8	81.1	79.5	87.8	79.6
+ FeedbackTrack	Ours	75.4	83.6	81.6	81.8	89.6	82.1
HIPTrack	CVPR'24	72.7	82.9	79.5	77.4	88.0	74.5
LoRAT-B_{378}	ECCV'24	72.9	81.9	79.1	73.7	82.6	72.9
LoRAT-L_{378}	ECCV'24	75.1	84.1	82.0	77.5	86.2	78.1
LoRAT-G_{378}	ECCV'24	76.2	85.3	83.5	78.9	87.8	80.7
SPMTrack-B	CVPR'25	74.9	84.0	81.7	76.5	85.9	76.3
+ FeedbackTrack	Ours	76.4	84.9	82.6	80.6	89.9	80.4
SPMTrack-L	CVPR'25	76.8	85.9	84.0	80.0	89.4	79.9
+ FeedbackTrack	Ours	<u>77.9</u>	<u>86.8</u>	<u>85.2</u>	<u>82.6</u>	<u>91.4</u>	<u>83.8</u>
SPMTrack-G	CVPR'25	77.4	86.6	85.0	81.0	89.2	82.3
+ FeedbackTrack	Ours	**79.1**	**87.8**	**86.3**	**83.4**	**91.8**	**84.0**



(a) ViT-B (b) ViT-L (c) ViT-G

Figure 4: Learned signed mean q-scale across feedback blocks, averaged over attention heads within each block. All q-scales are initialized to 0.01, as indicated by the dashed lines.

Analysis

Parameter efficiency. Table 5 reports the parameter overhead across both tracking frameworks. For the SPMTrack-based implementations, FeedbackTrack adds 0.983M, 2.620M, and 6.548M parameters to the B, L, and G variants,

Table 2: Ablation of Query Feedback (QF) and Gate Feedback (GF) on GOT-10k.

Backbone	QF	GF	AO	$SR_{0.5}$	$SR_{0.75}$
ViT-B			76.5	85.9	76.3
ViT-B	✓		79.7	89.2	80.1
ViT-B		✓	78.3	87.2	77.8
ViT-B	✓	✓	**80.6**	**89.9**	**80.4**
ViT-L			80.0	89.4	79.9
ViT-L	✓		82.0	91.1	83.4
ViT-L		✓	80.6	90.0	81.2
ViT-L	✓	✓	**82.6**	**91.4**	**83.8**
ViT-G			81.0	89.2	82.3
ViT-G	✓		83.0	91.4	83.7
ViT-G		✓	81.8	90.2	82.7
ViT-G	✓	✓	**83.4**	**91.8**	**84.0**

Table 3: Comparison between same-frame modulation and cross-frame feedback on GOT-10k. Same-frame modulation uses the detached current group input, while cross-frame feedback uses the detached previous-frame group output. Both variants use the same modules, insertion positions, and parameter count.

Backbone	Modulation Source	AO	$SR_{0.5}$	$SR_{0.75}$
ViT-B	None	76.5	85.9	76.3
ViT-B	Current-frame input	77.4	86.3	77.7
ViT-B	Previous-frame output	**80.6**	**89.9**	**80.4**
ViT-L	None	80.0	89.4	79.9
ViT-L	Current-frame input	80.3	90.2	81.1
ViT-L	Previous-frame output	**82.6**	**91.4**	**83.8**
ViT-G	None	81.0	89.2	82.3
ViT-G	Current-frame input	81.6	89.7	83.0
ViT-G	Previous-frame output	**83.4**	**91.8**	**84.0**

Table 4: Effect of RMS alignment on GOT-10k. All variants contain both Query Feedback and Gate Feedback.

Backbone	RMS Align.	AO	$SR_{0.5}$	$SR_{0.75}$
ViT-B		78.2	88.6	79.3
ViT-B	✓	**80.6**	**89.9**	**80.4**
ViT-L		82.1	90.3	82.6
ViT-L	✓	**82.6**	**91.4**	**83.8**
ViT-G		82.1	90.3	83.8
ViT-G	✓	**83.4**	**91.8**	**84.0**

corresponding to increases of only 0.852%, 0.690%, and 0.489%, respectively. For ARTrackV2-B and ARTrackV2-L, the corresponding increases are 0.983M and 2.620M, accounting for only 0.753% and 0.686% of the original model sizes. FeedbackTrack therefore introduces less than 1% parameter overhead across all five model configurations.

Runtime efficiency. Table 6 reports model inference throughput. For SPMTrack-B, SPMTrack-L, and SPMTrack-G, the throughput reductions are 7.44%, 5.81%, and 5.45%, respectively. For ARTrackV2-B and ARTrackV2-L, the reductions are only 0.18% and 0.86%. These results show that FeedbackTrack preserves most of the original inference throughput across both tracking frameworks while introducing explicit recurrent cross-frame modulation.

Table 5: Parameter overhead of FeedbackTrack across SPMTrack and ARTrackV2.

Model	Baseline	QF Only	QF + GF	Extra	Increase
SPMTrack-B	115.331M	115.409M	116.313M	0.983M	0.852%
SPMTrack-L	379.582M	379.791M	382.202M	2.620M	0.690%
SPMTrack-G	1339.512M	1340.034M	1346.060M	6.548M	0.489%
ARTrackV2-B_{384}	130.514M	130.593M	131.497M	0.983M	0.753%
ARTrackV2-L_{384}	382.059M	382.268M	384.679M	2.620M	0.686%

Table 6: Model inference throughput on the GOT-10k test set. Image loading and preprocessing are excluded.

Model	Baseline FPS	FeedbackTrack FPS	Reduction
SPMTrack-B	45.59	42.20	7.44%
SPMTrack-L	19.98	18.82	5.81%
SPMTrack-G	4.40	4.16	5.45%
ARTrackV2-B_{384}	15.7785	15.7505	0.18%
ARTrackV2-L_{384}	12.9309	12.8203	0.86%

Biologically inspired analysis. Biological visual feedback is distributed non-uniformly across the cortical hierarchy (Federer et al. 2021; Semedo et al. 2022; Fişek et al. 2023; Shen et al. 2022). We therefore analyze the signed mean q-scale across attention heads within each feedback block. Since all scales are initialized to 0.01, their depth-dependent variation is learned during optimization.

As shown in Fig. 4, ViT-B reaches its strongest feedback in the intermediate group, ViT-L develops stronger modulation from middle to deep groups, and ViT-G progressively increases toward deep groups before slightly decreasing. Overall, feedback is weaker in shallow groups and stronger in intermediate and deep groups. This pattern is qualitatively consistent with hierarchical biological feedback, but reflects computational correspondence rather than anatomical equivalence.

Conclusion

We presented FeedbackTrack, a visual-cortex-inspired framework that introduces sparse, group-level layer-aligned cross-frame feedback into Transformer trackers. Through complementary Query Feedback and Gate Feedback pathways, previous-frame representations recurrently modulate current visual processing while retaining the original feedforward pathway and tracking pipeline. Beyond visual tracking, the proposed mechanism provides a general approach for introducing recurrent temporal states into pretrained video models. Future work will explore its application to broader video tasks, including video object segmentation, action recognition, video understanding, and video generation.

References

Bai, Y.; Zhao, Z.; Gong, Y.; and Wei, X. 2024. ARTrackV2: Prompting Autoregressive Tracker Where to Concentrate. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19048–19057.

Cai, W.; Liu, Q.; and Wang, Y. 2024. HIPTrack: Visual Tracking with Historical Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19258–19267.

Cai, W.; Liu, Q.; and Wang, Y. 2025. SPMTrack: Spatio-Temporal Parameter-Efficient Fine-Tuning with Mixture of Experts for Scalable Visual Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16871–16881.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14572–14581.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8126–8135.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2022. MixFormer: End-to-End Tracking with Iterative Mixed Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13608–13618.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5374–5383.

Federer, F.; Ta'afua, S.; Merlin, S.; Hassanpour, M. S.; and Angelucci, A. 2021. Stream-Specific Feedback Inputs to the Primate Primary Visual Cortex. *Nature Communications*, 12(1): 228.

Fişek, M.; Herrmann, D.; Egea-Weiss, A.; Cloves, M.; Bauer, L.; Lee, T.-Y.; Russell, L. E.; and Häusser, M. 2023. Cortico-Cortical Feedback Engages Active Dendrites in Visual Cortex. *Nature*, 617(7962): 769–776.

Huang, L.; Zhao, X.; and Huang, K. 2019. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5): 1562–1577.

Lamme, V. A. F.; and Roelfsema, P. R. 2000. The Distinct Modes of Vision Offered by Feedforward and Recurrent Processing. *Trends in Neurosciences*, 23(11): 571–579.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4282–4291.

Markov, N. T.; Vezoli, J.; Chameau, P.; Falchier, A.; Quilodran, R.; Huissoud, C.; Lamy, C.; Misery, P.; Giroud, P.; Ullman, S.; Barone, P.; Dehay, C.; Knoblauch, K.; and Kennedy, H. 2014. Anatomy of Hierarchy: Feedforward and Feedback Pathways in Macaque Visual Cortex. *Journal of Comparative Neurology*, 522(1): 225–259.

Mayer, C.; Danelljan, M.; Paudel, D. P.; and Van Gool, L. 2021. Learning Target Candidate Association to Keep Track of What Not to Track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13444–13454.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Semedo, J. D.; Jasper, A. I.; Zandvakili, A.; Krishna, A.; Aschner, A.; Machens, C. K.; Kohn, A.; and Yu, B. M. 2022. Feedforward and Feedback Interactions between Visual Cortical Areas Use Different Population Activity Patterns. *Nature Communications*, 13(1): 1099.

Shen, S.; Jiang, X.; Scala, F.; Fu, J.; Fahey, P.; Kobak, D.; Tan, Z.; Zhou, N.; Reimer, J.; Sinz, F.; and Tolias, A. S. 2022. Distinct Organization of Two Cortico-Cortical Feedback Pathways. *Nature Communications*, 13(1): 6389.

Siu, C.; Balsor, J.; Merlin, S.; Federer, F.; and Angelucci, A. 2021. A Direct Interareal Feedback-to-Feedforward Circuit in Primate Visual Cortex. *Nature Communications*, 12(1): 4911.

Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021. Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1571–1580.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023. Autoregressive Visual Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9697–9706.

Xie, J.; Zhong, B.; Mo, Z.; Zhang, S.; Shi, L.; Song, S.; and Ji, R. 2024. Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19300–19309.

Yan, B.; Peng, H.; Fu, J.; Wang, D.; and Lu, H. 2021. Learning Spatio-Temporal Transformer for Visual Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10448–10457.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. In *Computer Vision – ECCV 2022*, volume 13682 of *Lecture Notes in Computer Science*, 341–357. Springer.

Zheng, Y.; Zhong, B.; Liang, Q.; Mo, Z.; Zhang, S.; and Li, X. 2024. ODTrack: Online Dense Temporal Token Learning for Visual Tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 7588–7596.