UNDERSTANDING AFFECTIVE ADAPTATION IN MULTIMODAL FOUNDATION MODELS: EMERGENT FUNCTIONAL SPECIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress in multimodal affective foundation models, how affective capabilities emerge within their internal architectures remains poorly understood. A critical open question is whether affective fine-tuning induces diffuse changes across the model or organizes computation into functionally specialized pathways. We systematically investigate this question through a broad module-level analysis of 13 affective model instances spanning nine model designs, multiple scales, tasks, and training paradigms, complemented by controlled functional analyses on representative models. We find that affective adaptation exhibits a consistent yet non-exclusive functional organization. Under matched trainable-parameter budgets, adapting only the feed-forward network (FFN) consistently outperforms adapting only the attention modules across all evaluated settings and, on average, nearly matches the performance obtained by tuning all major Transformer projections, identifying the FFN as a particularly efficient adaptation substrate. More strikingly, although the gate, up, and down projections exhibit comparable standalone adaptation capacity, their learned functional contributions become differentiated after joint optimization. Module recovery and targeted interventions identify `gate_proj` as a particularly prominent pathway, while checkpoint analysis shows that this differentiation develops over the course of training. We characterize this phenomenon as emergent functional specialization: distinct pathway-level roles are not fully explained by standalone adaptation capacity, but arise through joint affective adaptation. Building on this finding, Gate-Focused Efficient Tuning (GET) retains 96.2-98.0% of the performance obtained by tuning all major Transformer projections while using only 19.3-24.5% as many trainable parameters.

1 INTRODUCTION

Large language models and multimodal foundation models have demonstrated strong capabilities in affective understanding, including emotion recognition, sentiment analysis, affective conversation, and depression detection. Recent affective foundation models such as Emotion-LLaMA (Cheng et al., 2024), AffectGPT (Lian et al., 2025), OneEmo (Huang et al., 2026), and EAI-LLM (Lu et al., 2025) further show that general-purpose foundation models can be effectively adapted to a wide range of affect-related tasks across modalities. Despite this progress, considerably less is known about what happens *inside the model during affective adaptation*. In particular, it remains unclear how the acquired functionality is distributed across Transformer components and whether different computational pathways assume distinct roles through the adaptation process.

Recent mechanistic studies have shown that model capabilities can be associated with structured internal representations, including specific neurons, attention heads, pathways, and cross-layer circuits (Shao & Wu, 2025; Sofroniew et al., 2026; Wang et al., 2025). Such studies provide important evidence that learned capabilities need not be uniformly distributed throughout a model. However, identifying structures associated with an already acquired capability is different from understanding *how functional organization develops as that capability is acquired*. For affective models, this process remains largely unexplored. We therefore ask: **how does affective adaptation become functionally organized across Transformer pathways, and does this organization emerge through optimization?**

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.

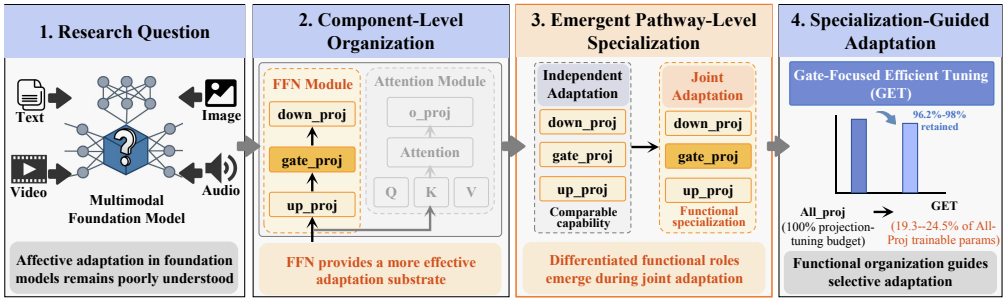


Figure 1: Overview of our study. We characterize affective adaptation from component-level organization to pathway-level functional specialization, examine how this specialization develops during joint optimization, and evaluate its practical utility for selective adaptation.

To answer this question, we conduct a broad analysis across 13 affective model instances spanning nine model designs, multiple backbone families, model scales, affective tasks, and adaptation paradigms. We use this broad model collection to characterize module-level adaptation patterns, and then conduct controlled pathway-level analyses on representative models to examine functional specialization more closely. We first examine adaptation at the level of major Transformer components. Parameter changes are strongly non-uniform, with FFN projections generally exhibiting larger adaptation-induced changes than Attention projections. Under comparable trainable-parameter budgets, adapting the FFN also tends to achieve stronger affective performance than adapting Attention. These observations suggest that, in the models studied here, the FFN may serve as a more effective substrate for affective adaptation.

We then move from component-level organization to individual FFN pathways. Although `gate_proj`, `up_proj`, and `down_proj` show broadly comparable standalone adaptation capacity, their functional contributions appear to diverge after joint optimization. Module recovery and targeted intervention results suggest that `gate_proj` becomes a particularly prominent pathway, while checkpoint analysis indicates that this differentiation emerges over the course of training. Taken together, these observations suggest that the resulting specialization cannot be fully explained by inherent differences in standalone adaptation capacity; instead, distinct functional roles may emerge through joint affective adaptation.

Finally, we examine whether the functional organization identified above can provide practical guidance for selective adaptation. As a simple mechanism-inspired validation, we restrict LoRA adaptation to the prominent gating pathway, leading to Gate-Focused Efficient Tuning (GET). Across AffectGPT, Emotion-LLaMA, and EAI-LLM, GET retains **96.2-98.0%** of the affective performance achieved by All-Proj while using only **19.3-24.5%** of its trainable parameters. Comparable performance retention is not observed in the examined non-affective controls. These results suggest that the functional organization revealed by our analysis may offer useful guidance for selective and parameter-efficient affective adaptation. Our main contributions are summarized as follows:

(1) We provide a systematic pathway-level analysis of affective adaptation in multimodal foundation models and show that it exhibits structured, non-uniform organization. Under comparable trainable-parameter budgets, the FFN supports stronger affective adaptation than Attention.

(2) We distinguish *standalone adaptation capacity* from *functional role under joint optimization*. Although multiple FFN pathways exhibit comparable standalone adaptation capacity, their functional contributions appear to diverge under joint optimization, with this differentiation emerging progressively over the course of training.

(3) We translate the observed specialization into Gate-Focused Efficient Tuning (GET), which retains 96.2–98.0% of All-Proj affective performance using only 19.3–24.5% of its trainable-parameter budget across the examined affective models.

2 Related Work

Affective Foundation Models. Large language models and multimodal foundation models have been increasingly adopted for affective computing tasks, including emotion recognition, sentiment

analysis, affective dialogue, and depression detection. Representative models such as Emotion-LLaMA (Cheng et al., 2024), AffectGPT (Lian et al., 2025), and EAI-LLM (Lu et al., 2025) demonstrate that large-scale foundation models can support affective understanding across language, speech, image, and video modalities. Recent studies further explore affective modeling under different supervision, domain, and multimodal settings (Huang et al., 2025; Chen et al., 2026; Song et al., 2025; Yang et al., 2025). Most existing work focuses on improving prediction accuracy, generation, robustness, or multimodal integration. In comparison, considerably less is known about how affective capabilities become functionally organized within the Transformer backbone during adaptation. Our work focuses on this internal organization rather than introducing a new affective task or model architecture.

Parameter-Efficient and Selective Adaptation. Foundation models are commonly adapted to downstream tasks through parameter-efficient fine-tuning (PEFT). LoRA (Hu et al., 2022), for example, represents task-specific updates using trainable low-rank factors while keeping pretrained parameters frozen. Subsequent work explores non-uniform allocation of adaptation capacity. AdaLoRA (Zhang et al., 2023) dynamically allocates rank budgets across weight matrices according to their estimated importance, while other studies investigate selective tuning over parameter subsets or Transformer layers (Ouyang et al., 2025; Zhou et al., 2025; Wei et al., 2025). These methods show that model components need not contribute equally to downstream adaptation. However, most selective PEFT approaches allocate capacity based on parameter importance, gradients, adapter outputs, layer selection, or downstream performance. Our work instead characterizes the capability-specific functional organization emerging during affective adaptation and examines whether it can guide target-module selection.

Mechanistic Analysis of Model Capabilities. Mechanistic studies have increasingly revealed structured internal organization underlying learned capabilities (Shao & Wu, 2025; Sofroniew et al., 2026; Wang et al., 2025). For example, Shao & Wu (2025) associate mathematical reasoning with particular Transformer pathways, including the attention output projection. In the affective domain, Sofroniew et al. (2026) study internal representations of emotion concepts, while Wang et al. (2025) identify emotion-related neurons, attention heads, and cross-layer circuits associated with emotional behavior. These studies mainly characterize structures associated with already acquired capabilities. In contrast, we study how differentiated functional roles develop across standard Transformer pathways during affective adaptation, distinguishing a pathway's standalone adaptation capacity from the role it assumes under joint optimization.

3 How Is Affective Adaptation Organized across Transformer Components?

3.1 Analyzed Transformer Components

We focus on the Transformer language backbones of the affective foundation models examined in this work. These models are typically adapted from general-purpose LLMs using parameter-efficient techniques such as LoRA, where the major linear projections in Attention and the FFN constitute common adaptation targets. Moreover, the Qwen- and LLaMA-based backbones examined in our experiments share a common projection structure, providing a consistent architectural basis for cross-model comparison.

As illustrated in Figure 2, we analyze seven projection pathways: `q_proj`, `k_proj`, `v_proj`, and `o_proj` in Attention, and `gate_proj`, `up_proj`, and `down_proj` in the gated FFN. These pathways allow us to compare adaptation both between the two major Transformer components (Attention versus FFN) and among their constituent projections.

Our analysis focuses on the shared Transformer backbone rather than attempting to characterize every modality-specific encoder or projector in the multimodal architecture.

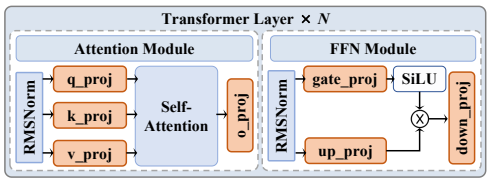
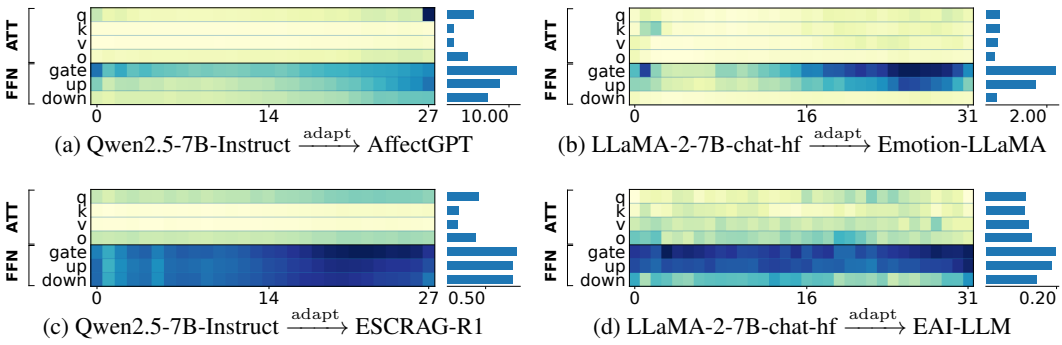


Figure 2: Schematic of the Transformer block analyzed in this work. Colored boxes denote the projection pathways considered in our module-level analysis.



(a) Qwen2.5-7B-Instruct $\xrightarrow{\text{adapt}}$ AffectGPT (b) LLaMA-2-7B-chat-hf $\xrightarrow{\text{adapt}}$ Emotion-LLaMA

(c) Qwen2.5-7B-Instruct $\xrightarrow{\text{adapt}}$ ESCRAG-R1 (d) LLaMA-2-7B-chat-hf $\xrightarrow{\text{adapt}}$ EAI-LLM

Figure 3: Module-wise parameter changes induced by affective adaptation across four representative models. Heatmaps show layer-wise Frobenius norms of weight differences, while bars report layer-averaged values. Across the examined models, the updates are non-uniformly distributed, with FFN projections generally exhibiting larger changes than Attention projections.

3.2 Affective Adaptation Produces Non-Uniform Parameter Changes

We first visualize four representative models—AffectGPT (Lian et al., 2025), Emotion-LLaMA (Cheng et al., 2024), ESCRAG-R1 (Liu et al., 2026), and EAI-LLM (Lu et al., 2025)—covering different backbone families, affective tasks, and adaptation settings. The broader analysis across 13 model instances is provided in Appendix A.1. For each projection X, we quantify the adaptation-induced parameter change using the Frobenius norm: $\Delta_X = \left\| W_X^{(\text{Affect})} - W_X^{(\text{Base})} \right\|_F$, where $W_X^{(\text{Affect})}$ and $W_X^{(\text{Base})}$ denote the weights of projection X in the adapted and base models, respectively. We compute Δ_X at each Transformer layer to characterize how parameter changes vary across projection pathways and network depth. As the raw Frobenius norm may also reflect differences in projection dimensionality, we treat it as a measure of overall weight change rather than per-parameter update magnitude.

As shown in Figure 3, the resulting changes are clearly non-uniform across Transformer components. Despite differences in backbone architecture and network depth, all four representative models exhibit a similar coarse-grained pattern: FFN projections generally undergo larger parameter changes than Attention projections. This consistency suggests that affective adaptation modifies different Transformer components to different extents rather than perturbing them uniformly.

We further assess robustness using complementary element-wise statistics and extend the analysis across additional model scales, architectures, tasks, and adaptation paradigms (Appendix A.1–A.4). The same coarse-grained asymmetry remains observable, suggesting that the pattern is not specific to the global Frobenius metric or the four representative models considered here. Importantly, however, larger parameter changes do not necessarily indicate greater functional relevance.

3.3 The FFN Provides a More Effective Adaptation Substrate

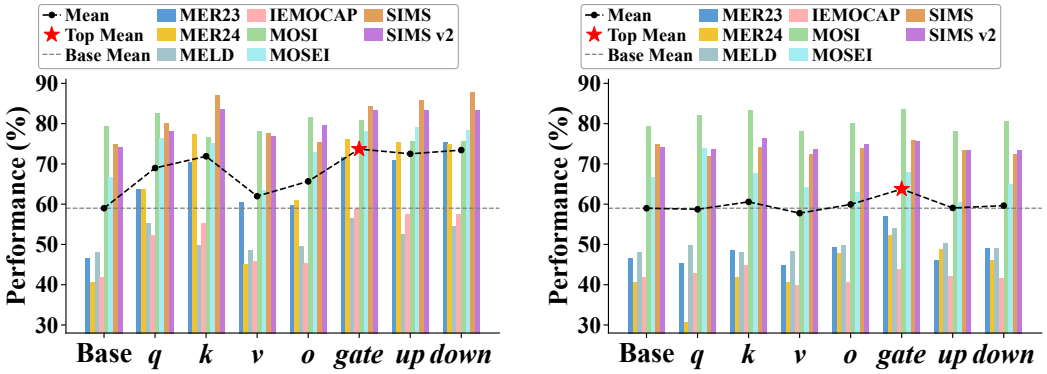
The parameter-space analysis above reveals systematically larger changes in the FFN, but update magnitude alone does not establish whether the FFN is functionally more effective for affective adaptation. We therefore compare Attention and FFN under controlled trainable-parameter budgets.

Specifically, we conduct matched-budget adaptation experiments on AffectGPT, Emotion-LLaMA, and EAI-LLM under three settings: Attention-only (ATT-Only), FFN-only (FFN-Only), and all projection modules (All-Proj). ATT-Only and FFN-Only use comparable numbers of trainable parameters, enabling a controlled comparison under similar adaptation budgets. Since the models employ different evaluation metrics, the Average row is reported only as an unweighted descriptive macro-average and should not be interpreted as a unified metric.

As shown in Table 1, FFN-Only consistently outperforms ATT-Only across all three models, yielding an average improvement of 1.33 points under comparable trainable-parameter budgets. At the same time, FFN-Only retains 99.2% of the All-Proj performance on average and even slightly ex-

Table 1: Performance of Attention-only and FFN-only adaptation under comparable trainable-parameter budgets, with All-Proj as the reference. Parentheses indicate trainable parameter counts and their percentages of the full model.

Model	ATT-only	FFN-only	$\Delta_{\mathrm{FFN-ATT}}$	All-Proj	FFN / All
AffectGPT	73.91 (10.09M, 0.13%)	**74.83** (9.46M, 0.12%)	**+0.92**	76.30	98.1%
Emotion-LLaMA	86.60 (93.32M, 1.37%)	**87.71** (92.79M, 1.36%)	**+1.11**	88.81	98.8%
EAI-LLM	71.78 (93.32M, 1.37%)	**73.73** (92.79M, 1.36%)	**+1.95**	73.01	101.0%
Average	77.43	**78.76**	**+1.33**	79.37	99.2%



(a) Standalone adaptation capacity

(b) Learned functional contribution

Figure 4: Standalone adaptation capacity and learned functional contribution across Transformer projections. (a) FFN pathways achieve comparable performance when optimized independently. (b) After joint adaptation, their functional roles differentiate, with `gate_proj` achieving the highest mean recovery performance.

ceeds All-Proj on EAI-LLM. Taken together, these results show that, under comparable trainable-parameter budgets, the FFN supports stronger affective adaptation than Attention while approaching the performance of All-Proj adaptation.

4 How Does Functional Specialization Arise Within the FFN?

Having established a component-level advantage of the FFN, we next examine how its internal pathways acquire distinct functional roles. We first compare their standalone adaptation capacity and then analyze how their contributions change under joint optimization.

4.1 Multiple FFN Pathways Exhibit Comparable Standalone Adaptation Capacity

We first ask whether functional specialization within the FFN simply reflects an intrinsic difference in standalone adaptation capacity. To examine this question, we independently optimize each projection while keeping all other model parameters frozen. This includes the four Attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`) and the three FFN projections (`gate_proj`, `up_proj`, and `down_proj`). The full set of projections provides a common reference across the Transformer, while our main focus is whether the three FFN pathways differ substantially in their standalone affective adaptation capacity.

As shown in Figure 4a, all three FFN pathways achieve strong affective performance when optimized independently and outperform most Attention projections. More importantly, their average performance remains relatively close. Although minor differences exist, no single FFN pathway exhibits an overwhelming advantage under independent adaptation, indicating broadly comparable standalone adaptation capacity among the three pathways.

Table 2: Targeted interventions on projection-wise LoRA updates in AffectGPT and OneEmo. Values in parentheses denote absolute performance changes relative to the All-Proj baseline.

Model	Module	Removal	Randomization	Direction Reversal
AffectGPT	All-Proj	76.30	76.30	76.30
	q_proj	68.16 (↓8.14)	67.67 (↓8.63)	65.20 (↓11.10)
	k_proj	68.78 (↓7.52)	68.46 (↓7.84)	69.39 (↓6.91)
	v_proj	69.07 (↓7.23)	69.25 (↓7.05)	69.05 (↓7.25)
	o_proj	69.69 (↓6.61)	68.94 (↓7.36)	69.63 (↓6.67)
	gate_proj	**65.66 (↓10.64)**	**66.20 (↓10.10)**	**55.97 (↓20.33)**
	up_proj	68.99 (↓7.31)	68.88 (↓7.42)	67.03 (↓9.27)
	down_proj	69.04 (↓7.26)	69.63 (↓6.67)	68.35 (↓7.95)
OneEmo	All-Proj	71.80	71.80	71.80
	q_proj	69.95 (↓1.85)	70.12 (↓1.68)	68.15 (↓3.65)
	k_proj	71.70 (↓0.10)	72.07 (↑0.27)	72.00 (↑0.20)
	v_proj	72.48 (↑0.68)	71.54 (↓0.26)	70.87 (↓0.93)
	o_proj	71.00 (↓0.80)	71.43 (↓0.37)	71.88 (↑0.08)
	gate_proj	**66.33 (↓5.47)**	**66.14 (↓5.66)**	**62.80 (↓9.00)**
	up_proj	71.62 (↓0.18)	71.47 (↓0.33)	68.50 (↓3.30)
	down_proj	71.31 (↓0.49)	70.86 (↓0.94)	70.00 (↓1.80)

4.2 Jointly Adapted Pathways Exhibit Differentiated Functional Contributions

We next examine the functional roles assumed by individual pathways after joint adaptation. Starting from the fully adapted All-Proj model, we separately recover the learned LoRA updates of each projection in the corresponding base model, while leaving all other adaptation updates unloaded. The resulting performance measures how much of the jointly learned affective functionality can be recovered from each pathway in isolation.

As shown in Figure 4b, the pattern differs markedly from standalone adaptation. While the three FFN pathways exhibit broadly comparable adaptation capacity when optimized independently, their learned contributions become clearly differentiated after joint optimization. In particular, recovering only the gate_proj updates yields the highest mean performance, whereas up_proj and down_proj recover less.

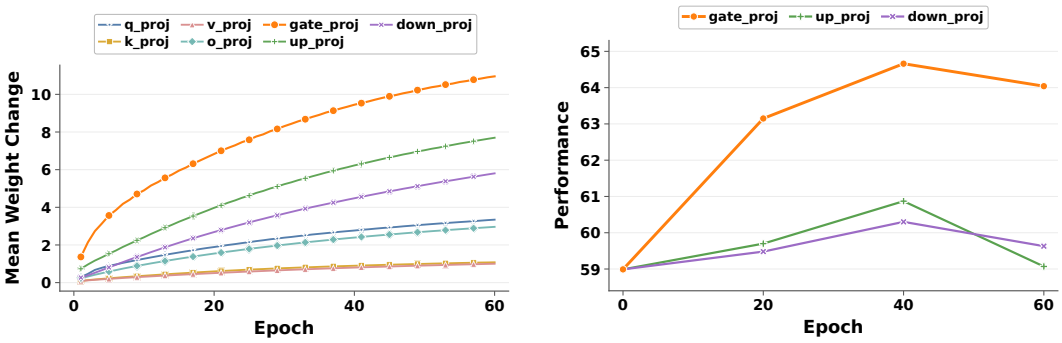
This contrast shows that comparable standalone adaptation capacity does not necessarily translate into comparable functional roles after joint optimization. Instead, the jointly adapted model develops an asymmetric distribution of learned functionality across FFN pathways, with gate_proj emerging as the most prominent pathway in this recovery analysis.

4.3 Targeted Interventions Reveal Uneven Functional Dependency

The recovery analysis above characterizes how much jointly learned functionality can be recovered from each pathway in isolation, but it does not directly measure how strongly the fully adapted model depends on that pathway. We therefore complement the recovery analysis with targeted interventions on the learned projection-wise LoRA updates.

For each projection pathway, we apply three interventions to its learned update ΔW. **Removal** sets the corresponding update to zero; **Randomization** replaces the learned update with randomized values; and **Direction Reversal** replaces ΔW with $-\Delta W$. All other learned updates remain unchanged. We then measure the resulting degradation in affective performance relative to the original All-Proj model.

As shown in Table 2, intervention sensitivity varies substantially across projection pathways, with a consistent pattern across both models. In AffectGPT, disrupting gate_proj produces the largest performance degradation under all three interventions. The effect is particularly pronounced under



(a) Evolution of module-wise LoRA updates

(b) Epoch-wise recovery of FFN pathways

Figure 5: Training dynamics of pathway differentiation during affective adaptation. (a) LoRA update magnitudes across Transformer projections become increasingly differentiated over training. (b) Checkpoint-wise recovery of the three FFN pathways reveals emerging functional differentiation, with `gate_proj` becoming more prominent during training.

Direction Reversal, where replacing its learned update ΔW with $-\Delta W$ reduces performance from 76.30 to 55.97, a drop of 20.33 points.

A similar pattern emerges in OneEmo. Removing or randomizing the `gate_proj` update reduces performance by 5.47 and 5.66 points, respectively, substantially more than interventions on the other projections. Direction Reversal again results in the largest performance degradation for `gate_proj`, causing a 9.00-point drop. In contrast, interventions on several Attention projections have relatively small effects and occasionally yield slight performance improvements. These results indicate that the stronger dependency on the gating pathway is not specific to AffectGPT.

The intervention results provide evidence complementary to the recovery analysis. Whereas recovery measures how much jointly learned functionality can be recovered from a pathway in isolation, intervention measures how sensitive the fully adapted model is to disrupting that pathway. Together, these analyses provide converging evidence that joint affective adaptation produces differentiated functional roles across otherwise capable FFN pathways, with `gate_proj` emerging as a particularly prominent contributor.

4.4 Functional Differentiation Develops During Affective Adaptation

The preceding analyses characterize differentiated functional roles among FFN pathways in the final adapted model. We next examine how these differences develop during joint optimization by tracking parameter changes across Transformer projections and functional recovery of the three FFN pathways throughout training.

We first examine the evolution of pathway-wise parameter updates. As shown in Figure 5a, the learned LoRA update magnitudes become increasingly differentiated over training. In this trajectory, `gate_proj` accumulates the largest changes, while the other projections follow distinct optimization trajectories. A similar differentiation of parameter updates is observed in an additional model (Appendix A.5), although the relative ordering among FFN pathways varies across models.

We then ask whether this parameter-level differentiation is accompanied by changes in functional contribution. For `gate_proj`, `up_proj`, and `down_proj`, we recover their learned LoRA updates from intermediate checkpoints and evaluate each pathway in isolation. As shown in Figure 5b, the three pathways exhibit similar recovery performance at early stages, but their contributions diverge as optimization proceeds. By intermediate epochs, `gate_proj` recovers noticeably higher performance than `up_proj` and `down_proj`. Although the gap narrows toward the final checkpoint, `gate_proj` remains the strongest recovered FFN pathway.

These trajectories show that the functional differentiation observed after joint adaptation is not apparent at early stages but develops during optimization. Combined with the comparable standalone

Table 3: Performance retention and trainable-parameter efficiency of GET across affective and non-affective settings. Param. Ratio is relative to All-Proj, and Retention is the GET-to-All-Proj performance ratio.

Setting	Model	Evaluation	Param. Ratio	Retention	GET / All-Proj
Affective Understanding	AffectGPT	8 datasets (avg.)	24.5%	**96.6%**	73.68 / 76.30
	Emotion-LLaMA	1 dataset	19.3%	**98.0%**	87.03 / 88.81
	EAI-LLM	3 datasets (avg.)	19.3%	**96.2%**	70.25 / 73.01
Mathematical Reasoning	LLaMA (Gaussian)	2 datasets (avg.)	19.3%	79.7%	19.27 / 24.18
Instruction Following	LLaMA (QLoRA)	1 dataset	19.3%	74.0%	32.93 / 44.50

capacity observed in Section 4.1, this suggests the distinct roles of FFN pathways are shaped through joint adaptation rather than simply inherited from differences in their individual adaptation capacity.

5 From Emergent Specialization to Efficient Affective Adaptation

The preceding analyses reveal that joint affective adaptation develops differentiated functional roles across FFN pathways, with `gate_proj` exhibiting a particularly prominent contribution. We next examine whether this empirical specialization can provide useful guidance for parameter-efficient affective adaptation. Rather than treating the gating pathway as an inherently optimal or predefined affective module, we use the observed functional organization as a basis for selective adaptation.

5.1 Gate-Focused Efficient Tuning

Based on this observation, we introduce **Gate-Focused Efficient Tuning (GET)**, a simple selective adaptation strategy that restricts trainable LoRA updates to the FFN gating pathway. Specifically, GET attaches LoRA adapters only to `gate_proj` in each Transformer layer, while the Attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`) and the remaining FFN projections (`up_proj` and `down_proj`) receive no trainable adapters. During affective adaptation, only the LoRA parameters associated with `gate_proj` are optimized.

GET therefore serves as a direct test of whether the functional specialization identified in the preceding analyses can be translated into a substantially smaller adaptation budget while preserving affective performance.

5.2 Performance Retention and Non-Affective Controls

We evaluate whether the specialization-guided selection used by GET can preserve affective performance under a substantially reduced trainable parameter budget. For each model, GET is compared with All-Proj adaptation, and we report both the fraction of trainable parameters used by GET and its performance retention relative to the corresponding All-Proj baseline.

For the affective settings, we follow the original fine-tuning and evaluation protocol of each model. AffectGPT is evaluated by averaging performance across eight affective benchmarks, Emotion-LLaMA is evaluated on its original benchmark, and EAI-LLM is evaluated by averaging performance across three bodily emotion recognition datasets. Dataset descriptions and evaluation metrics are provided in Appendix A.7 and Appendix A.8, respectively. Because the models use different tasks and evaluation metrics, absolute scores are not compared across models; instead, performance retention is computed within each model relative to its own All-Proj baseline.

As shown in Table 3, GET retains 96.2–98.0% of the corresponding All-Proj performance across the three affective models while using only 19.3–24.5% of the All-Proj trainable-parameter budget. This consistent within-model retention indicates that the functional specialization identified above

can provide a useful basis for substantially reducing the adaptation budget without sacrificing most of the acquired affective performance.

We additionally evaluate the same gate-focused strategy in two non-affective control settings, mathematical reasoning and instruction following, following the experimental protocols of (Meng et al., 2024; Dettmers et al., 2023), respectively. Under a comparable 19.3% parameter ratio, GET retains only 79.7% and 74.0% of the corresponding All-Proj performance, respectively. Thus, the high retention observed in the affective models is not uniformly reproduced in these non-affective controls. While these comparisons are not sufficient to establish that gate-focused adaptation is uniquely specific to affective capabilities, they provide a useful boundary test against the alternative explanation that restricting adaptation to `gate_proj` is generally effective across downstream tasks.

The above results show that the pathway specialization uncovered by our analysis has practical value: using `gate_proj` as the sole adaptation target preserves most of the All-Proj performance across the examined affective models with a substantially smaller trainable-parameter budget.

6 Discussion

Interpretation of Emergent Functional Specialization. Our results highlight a distinction between a pathway's *capacity* to support affective adaptation and the *functional role* it assumes under joint optimization. Although the FFN pathways exhibit broadly comparable standalone adaptation capacity, their contributions become differentiated when optimized together. The prominence of `gate_proj` should therefore not be interpreted as evidence for a predefined "emotion module," but rather as an emergent property of joint adaptation.

This perspective suggests that understanding post-training requires not only asking which components can acquire a capability, but also how joint optimization distributes learned functionality across interacting pathways. The functional organization emerging after adaptation need not be predictable from the standalone capacity of individual components.

The computational role of gating offers one possible interpretation. Because gated FFNs multiplicatively modulate intermediate features, affective adaptation may benefit from selectively emphasizing or suppressing existing representations rather than uniformly rewriting them. This view is conceptually consistent with accounts of affective modulation in neuroscience (Anderson & Phelps, 2001; Janak & Tye, 2015). However, this remains a hypothesis: our experiments do not establish selective modulation as the underlying mechanism, nor imply a direct correspondence between Transformer gating and biological affective processing.

Scope and Limitations. Our analysis operates at the level of Transformer projection pathways and does not yet explain affective processing at the activation, feature, or circuit level. Moreover, the prominence of `gate_proj` should not be assumed to generalize universally across architectures, tasks, or post-training procedures. Broader model families and finer-grained mechanistic analyses will be needed to determine the generality and underlying mechanism of the observed specialization.

7 Conclusion

This work investigates how affective adaptation becomes functionally organized within multimodal foundation models. Across the examined settings, the FFN supports stronger affective adaptation than Attention under comparable trainable-parameter budgets, while its internal pathways exhibit comparable standalone capacity but differentiated functional contributions after joint optimization. Training dynamics further show that this differentiation develops during adaptation, with `gate_proj` emerging as a particularly prominent pathway. Building on this organization, GET retains 96.2–98.0% of the corresponding All-Proj affective performance using only 19.3–24.5% of its trainable-parameter budget.

Overall, our findings suggest that affective capability is not simply localized in a fixed architectural module, but can become *functionally organized through adaptation*. More broadly, we hope this perspective encourages a shift from asking only *where* a learned capability resides to understanding *how* its functional organization emerges during learning, providing a starting point for future studies of capability formation, interpretation, and selective adaptation in foundation models.

AI USE STATEMENT

Generative AI tools were used to assist with language compression and LaTeX formatting. The authors remain responsible for verifying all technical content, citations, claims, and reported results.

ETHICS STATEMENT

All emotion-related datasets used in this work are publicly available research datasets and are used solely for academic research purposes. We do not collect new data from human participants or access private or personally identifiable information beyond what is included in the publicly released datasets. Our experiments focus on analyzing model adaptation and do not involve direct interactions with human subjects or deployment in real-world decision-making settings. We use all datasets in accordance with their intended research purposes and applicable licenses or terms of use. To the best of our knowledge, this work does not introduce additional ethical concerns beyond those associated with the use of existing public datasets. The authors declare no conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide our code in an anonymous repository at anonymous repository. Sections 3–5 examine the allocation and evolution of adaptation pathways in affective LLMs from complementary perspectives. The appendix presents additional experiments that complement the main findings, together with detailed descriptions of the datasets, models, and evaluation metrics used in our experiments.

REFERENCES

Adam K Anderson and Elizabeth A Phelps. Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature*, 411(6835):305–309, 2001.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246. Association for Computational Linguistics, July 2018.

Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

Jiuyi Chen, Mingkui Tan, Haifeng Lu, Qiuna Xu, Zhihua Wang, Runhao Zeng, and Xiping Hu. Towards stable cross-domain depression recognition under missing modalities. *Pattern Recognition*, pp. 113367, 2026.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Nesrine Fourati and Catherine Pelachaud. Perception of emotions and body movement in the emilya database. *IEEE Transactions on Affective Computing*, 9(1):90–101, Jan 2018. doi: 10.1109/taffc.2016.2591039.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

Jiahao Huang, Zheng Lian, Jingyi Zhang, Zhide Chen, Xiaojiang Peng, and Shaonan Wang. Oneemo: A unified multimodal reasoning model for emotion perception, understanding, and interaction. *arXiv preprint arXiv:2608.06013*, 2026.

Xuandong Huang, Yuzhe Zhou, Jiashu Li, Shiqian Lu, and Shangfei Wang. Emodetective: Detecting, exploring, and thinking emotional cause in videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5735–5744, 2025.

Patricia H Janak and Kay M Tye. From circuits to behaviour in the amygdala. *Nature*, 517(7534): 284–292, 2015.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9610–9614. Association for Computing Machinery, 2023. ISBN 9798400701085.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48. Association for Computing Machinery, 2024.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

Weichu Liu, Yuxuan Hu, Yirong Sun, Ningning Mao, Ziyun Zhang, Jian Chen, Mingyang Xu, Qishan Zhong, and Chengming Li. Escrag-r1: Retrieval-augmented reinforcement learning for emotional support conversation. *arXiv preprint arXiv:2608.21925*, 2026.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module, 2022.

Haifeng Lu, Jiuyi Chen, Feng Liang, Mingkui Tan, Runhao Zeng, and Xiping Hu. Understanding emotional body expressions via large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1447–1455, 2025.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Ziheng Ouyang, Zhen Li, and Qibin Hou. K-lora: Unlocking training-free fusion of any subject and style loras. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13041–13050. IEEE, 2025.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883, 2017.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536. Association for Computational Linguistics, July 2019.

Tomasz Sapiński, Dorota Kamińska, Adam Pelikant, Cagri Ozcinar, Egils Avots, and Gholamreza Anbarjafari. Multimodal database of emotional speech, video and gestures. In *Pattern Recognition and Information Forensics: ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers 24*, pp. 153–163. Springer, 2019.

Jie Shao and Jianxin Wu. Who reasons in the large language models? *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, et al. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026.

Changhao Song, Yazhou Zhang, Hui Gao, Kaiyun Huang, and Peng Zhang. Emotion-o1: Adaptive long reasoning for emotion understanding in llms. *arXiv preprint arXiv:2505.22548*, 2025.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, et al. Do llms" feel"? emotion circuits discovery and control. *arXiv preprint arXiv:2510.11328*, 2025.

Chenxing Wei, Yao Shu, Ying Tiffany He, and Fei Yu. Flexora: Flexible low-rank adaptation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 14643–14682, 2025.

Minqiang Yang, Zhichao Yang, Zhaolong Ning, Hao Shen, Chengsheng Mao, Changsheng Ma, and Bin Hu. Echoes of empathy: a symbiotic iot-based emotion feedback framework for psychological interventions via large language model. *IEEE Internet of Things Journal*, 12(20): 41297–41308, 2025.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727. Association for Computational Linguistics, July 2020.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114. Association for Computational Linguistics, September 2017.

Mingming Zhang, Lu Yu, Keye Zhang, Bixuan Du, Bin Zhan, Shaohua Chen, Xiuhao Jiang, Shuai Guo, Jiafeng Zhao, Yang Wang, Bin Wang, Shenglan Liu, and Wenbo Luo. Kinematic dataset of actors expressing emotions. *Scientific Data*, Sep 2020. doi: 10.1038/s41597-020-00635-7.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao, and Muyun Yang. Lora-drop: Efficient lora parameter pruning based on output evaluation. In *Proceedings of the 31st international conference on computational linguistics*, pp. 5530–5543, 2025.

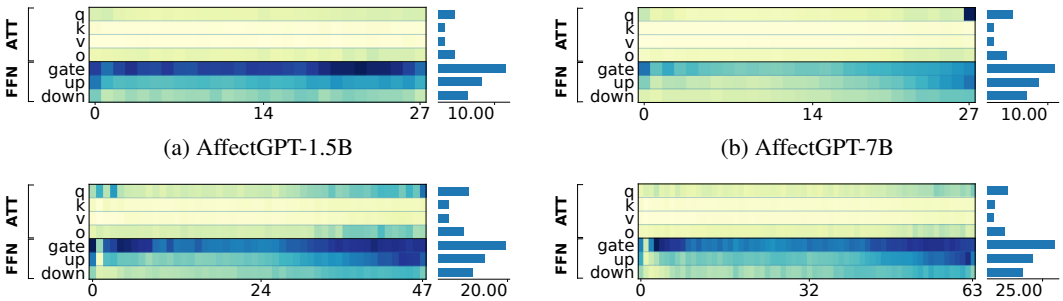
A APPENDIX

APPENDIX CONTENTS

A.1 Module-Level Adaptation Patterns across Model Families and Scales 13
A.2 Generality across Affective Tasks and Training Paradigms 14
A.3 Robustness Analysis with Alternative Weight-Change Metrics 15
A.4 Comparison with Non-Affective Adaptation 16
A.5 Training Dynamics across Additional Models 17
A.6 Additional Analysis of Module-wise Functional Contributions 18
A.6.1 Compositional Analysis of Recovered LoRA Updates 18
A.6.2 Controlled Analysis of Module Adaptation Capacity 19
A.7 Downstream Dataset Overview 20
A.8 Performance Metrics 22

A.1 MODULE-LEVEL ADAPTATION PATTERNS ACROSS MODEL FAMILIES AND SCALES

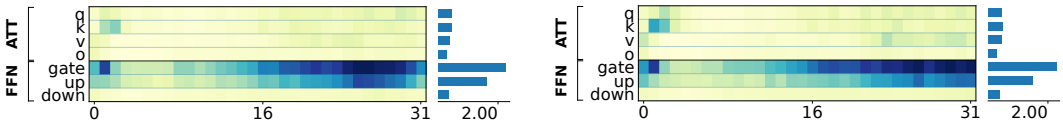
To examine whether the observed module-level adaptation pattern generalizes across model families and scales, we extend our analysis to multiple model sizes within both the AffectGPT and Emotion-LLaMA series. Specifically, we analyze Qwen2.5-based AffectGPT models at 1.5B, 7B, 14B, and 32B scales, together with LLaMA-2-based Emotion-LLaMA models at 7B and 13B scales. For each model, we quantify module-wise parameter changes using a global Frobenius metric, computed between the projection weights of the affectively adapted model and those of its corresponding base model across all layers.



(a) AffectGPT-1.5B (b) AffectGPT-7B

(c) AffectGPT-14B (d) AffectGPT-32B

Figure 6: Layer-wise module weight changes across model families and scales. Across AffectGPT models of different sizes, affective adaptation consistently exhibits an FFN-dominant, non-uniform pattern of parameter changes.



(a) Emotion-LLaMA-7B (b) Emotion-LLaMA-13B

Figure 7: Layer-wise module weight changes across model families and scales. Across Emotion-LLaMA models of different sizes, affective adaptation consistently exhibits an FFN-dominant, non-uniform pattern of parameter changes.

As shown in Figure 6 and 7, a consistent module-level pattern emerges across both model families and scales. Attention projections generally exhibit relatively limited parameter changes, whereas

Table 4: Affective models included in the extended generality analysis. The models cover diverse affective tasks and adaptation paradigms.

Model	Affective Task	Training Strategy
LoRA-based fine-tuning		
AffectGPT	Open-vocabulary emotion recognition and emotion description	LoRA
Emotion-LLaMA	Emotion recognition	LoRA
SCD-MLLM	Multimodal depression recognition	LoRA
EAI-LLM	Skeleton-based emotion recognition	LoRA
EFT-CoT	Emotion-focused therapy and therapeutic-process reasoning	LoRA
llama-3.1-8b-financial-sentiment	Three-way financial sentiment classification	QLoRA
Other adaptation paradigms		
ESCRAG-R1	Emotional support dialogue	SFT + GRPO
EmotionThinker	Speech emotion recognition	SFT + LoRA + GRPO
VidEmo-T1	Fine-grained video emotion and facial-expression analysis	SFT + GRPO

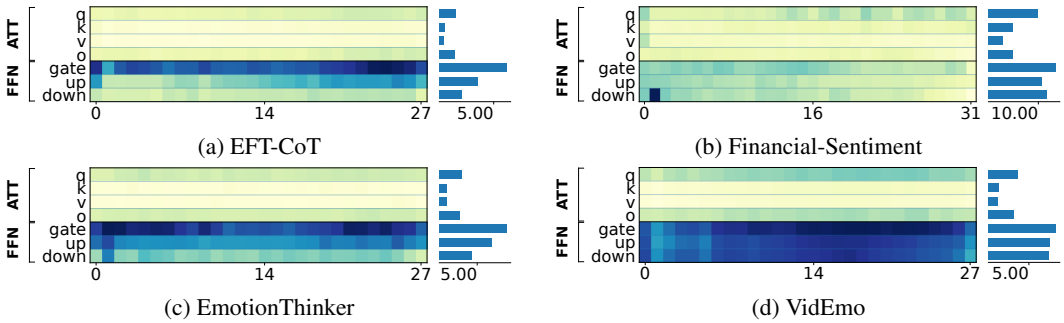
FFN projections undergo substantially more pronounced changes throughout the network. This FFN-dominant pattern is evident from the smallest AffectGPT-1.5B model to the larger 14B and 32B variants, and is similarly observed in both Emotion-LLaMA-7B and Emotion-LLaMA-13B. At the same time, the relative magnitude and layer-wise distribution of changes among individual FFN projections vary across models, indicating that the detailed adaptation pattern is not tied to a particular projection. These results show that the non-uniform organization of affective adaptation is robust across model families and scales. Despite differences in architecture and parameter size, affective adaptation consistently induces stronger changes in the FFN than in Attention, supporting **the generality of the FFN-dominant module-level adaptation pattern rather than a scale- or architecture-specific effect**.

A.2 Generality across Affective Tasks and Training Paradigms

To further examine the generality of the observed module-level adaptation pattern, we extend our analysis to a broader collection of affective models covering diverse tasks and training paradigms. As summarized in Table 4, these models span open-vocabulary emotion recognition, emotion description, emotional support dialogue, multimodal depression recognition, skeleton-based emotion recognition, speech emotion recognition, video emotion analysis, sentiment analysis, empathetic response generation, and therapy-oriented affective reasoning. Their adaptation strategies also vary substantially, including LoRA-based fine-tuning, reinforcement learning, GRPO-based optimization, and full-parameter supervised fine-tuning. We further provide visualizations of the additional representative affective models not shown in the main text, as shown in Figure 8, providing additional evidence that the observed module-level adaptation pattern is consistent across diverse model architectures and training paradigms.

For each model, we compare the adapted model with its corresponding base model and analyze module-wise parameter changes across Attention and FFN projections following the same protocol used in our main analysis. Since these models differ substantially in architecture, parameter scale, task formulation, and training procedure, we focus on the relative distribution of parameter changes within each model rather than directly comparing their absolute magnitudes.

Across this diverse set of affective models, we consistently observe non-uniform module-level adaptation patterns. Despite substantial differences in affective task formulation and optimization strategy, FFN projections generally exhibit more pronounced changes than Attention projections. This trend is observed not only in conventional emotion recognition and sentiment analysis, but also in affective generation, multimodal depression recognition, speech- and video-based emotion understanding, and affective reasoning tasks.



(a) EFT-CoT (b) Financial-Sentiment

(c) EmotionThinker (d) VidEmo

Figure 8: Visualizations of the affective models not shown in the main text.

These results provide additional evidence that the FFN-dominant organization observed in our main experiments is not confined to a particular affective task, modality, or training paradigm. **Across the examined affective settings, structured and non-uniform parameter changes emerge as a recurring pattern**.

A.3 Robustness Analysis with Alternative Weight-Change Metrics

To further verify that the observed parameter changes are not specific to a particular distance metric, we evaluate the magnitude of weight updates using several complementary statistics. For each module, we define the parameter update as $\Delta W = W_{\text{aff}} - W_{\text{base}}$, where W_{aff} and W_{base} denote the weights of the affectively adapted and base models, respectively. Specifically, we consider the mean absolute update,

$$\text{MAE}(\Delta W) = \frac{1}{N}\sum_{i=1}^{N} |\Delta w_i|,$$

the median absolute update,

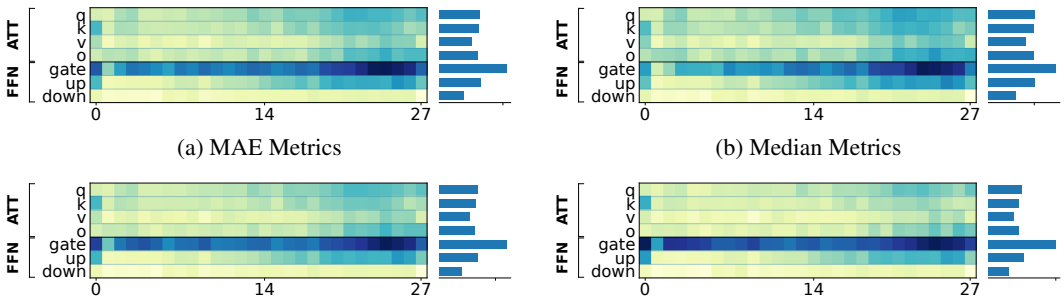
$$\text{Median}(|\Delta W|),$$

and the 90th, and 99th percentiles of the absolute update distribution,

$$P_{90}(|\Delta W|), \quad P_{99}(|\Delta W|).$$

These metrics capture complementary aspects of the parameter-update distribution. MAE measures the average magnitude of parameter changes, while the median reflects the typical update magnitude and is less sensitive to extreme values. The upper percentiles characterize relatively large updates and provide a more robust view of the upper tail of the distribution without relying on a small number of extreme values.

Across these alternative metrics, a consistent module-wise asymmetry remains evident, as shown in Figure 9. In particular, `gate_proj` exhibits the most pronounced parameter changes across MAE, median, P_{90}, and P_{99}, while `up_proj` also shows relatively large updates compared with most Attention projections. Importantly, this pattern appears not only in the upper tail of the update distribution but also in its mean and median. The consistency across multiple distributional statistics suggests that the prominence of `gate_proj` is not driven solely by a small number of extreme parameter updates, but instead reflects a broader shift in its parameter distribution during affective adaptation. These results further support the robustness of the non-uniform adaptation pattern observed across Transformer modules.



(a) MAE Metrics (b) Median Metrics

(c) P_{90} Metrics (d) P_{99} Metrics

Figure 9: Robustness of module-wise parameter changes under alternative weight-change metrics on EFT-CoT. Panels (a)–(d) report layer-wise parameter changes measured by MAE, median absolute update, P_{90}, and P_{99}, respectively. The bar chart in each panel reports the corresponding metric aggregated across layers. Across these metrics, FFN projections generally exhibit larger parameter changes than Attention projections, indicating that the observed module-wise asymmetry is robust to the choice of weight-change metric.

Table 5: Non-affective models included in the extended generality analysis. The models cover diverse tasks and post-training paradigms.

Model	**Task**	**Training Strategy**
Instruction fine-tuning		
deepseek-coder-1.3b -instruct	Code generation, code completion, programming question answering, and instruction following	Instruction SFT
SmolLM-1.7B-Instruct (v0.2)	English general question answering, daily dialogue, creative writing, and basic Python programming	SFT
Reasoning and capability distillation		
DeepSeek-R1-Distill -Qwen-1.5B/7B/14B	Mathematical, coding, and general reasoning with long reasoning chains	Knowledge distillation via SFT
Qwen3-0.6B	Controllable thinking and non-thinking, mathematical, coding, and logical reasoning, instruction following, tool use, and multilingual generation	Strong-to-weak distillation (off-policy $\rightarrow$ on-policy)
Preference optimization and reinforcement learning		
Qwen2.5-Coder-1.5B -Instruct	Code generation, code reasoning, code repair, and programming instruction following	SFT $\rightarrow$ DPO
granite-3.2-2b -instruct	Controllable reasoning, instruction following, question answering, summarization, RAG, coding, and tool use	TPO-based RL
OLMo-2-0425-1B-DPO	General dialogue, instruction following, response-preference alignment, and mathematical tasks	DPO
OLMo-2-0425-1B -RLVR1	Mathematical reasoning and verifiable instruction following under mixed GSM, MATH, and IF constraints	RLVR

A.4 Comparison with Non-Affective Adaptation

To provide an additional comparison, we extend the module-level analysis to several non-affective models spanning question answering, code generation, mathematical reasoning, and logical reasoning, as shown in Tab 5. These models cover different post-training paradigms, including instruction fine-tuning, knowledge distillation, preference optimization, and reinforcement learning. We apply the same module-wise analysis to examine how parameter changes are distributed across Attention and FFN projections.

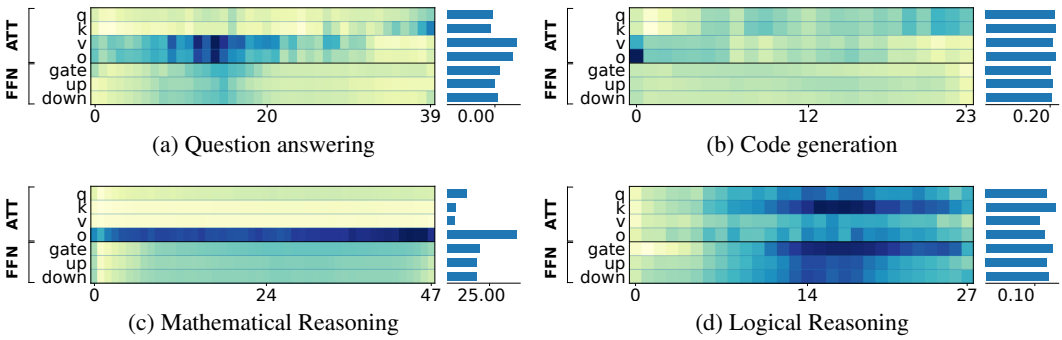


Figure 10: Module-level parameter changes across non-affective tasks, including (a) question answering (granite-3.2-2b-instruct), (b) code generation (deepseek-coder-1.3b-instruct), (c) mathematical reasoning (DeepSeek-R1-Distill-Qwen-14B), and (d) logical reasoning (Qwen3-06B). The heatmaps show layer-wise parameter changes across Transformer projections, with the horizontal axis indicating layer indices; the bars on the right report the mean L2 difference of each projection across layers. The examined tasks exhibit diverse module-level patterns: question answering and mathematical reasoning show relatively pronounced changes in `o_proj`, whereas code generation and logical reasoning display more distributed changes across projections and layers.

As shown in Figure 10, the examined non-affective settings exhibit diverse module-level patterns. For example, question answering and mathematical reasoning show relatively pronounced changes in `o_proj`, whereas code generation and logical reasoning display more distributed changes across multiple projections and layers. Overall, both the dominant projections and their layer-wise distributions vary across the examined settings. These results indicate that the module-level patterns observed in the non-affective models differ from those identified in the affective models. In particular, the consistent FFN-dominant tendency observed across the examined affective settings is not reproduced uniformly in these non-affective cases. Although the current comparisons are insufficient to establish task-specific functional organization, they provide additional evidence that the observed FFN-dominant pattern is not simply shared across all forms of post-training.

A.5 Training Dynamics across Additional Models

To examine whether the parameter-level differentiation observed during training is specific to AffectGPT, we additionally track projection-wise LoRA update magnitudes in EAI-LLM (Lu et al., 2025). As shown in Figure 11, the pathways again exhibit clearly non-uniform optimization trajectories. In this model, `gate_proj` and `up_proj` show similarly large updates, while `down_proj` and the Attention projections follow distinct trajectories. These results suggest that pathway-level differentiation during adaptation is not limited to a single model, although the relative ordering among FFN pathways may vary across architectures.

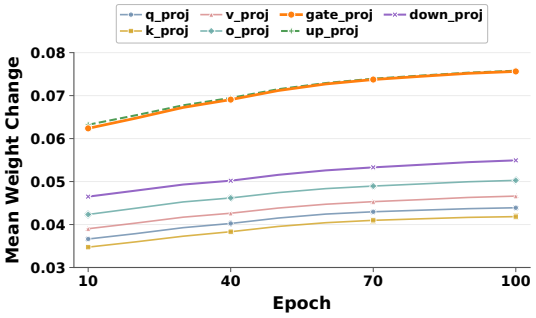


Figure 11: Training dynamics of projection-wise LoRA update magnitudes in EAI-LLM. Different projection pathways exhibit non-uniform optimization trajectories, with `gate_proj` and `up_proj` showing similarly large updates, while `down_proj` and the Attention projections follow distinct trajectories.

A.6 Additional Analysis of Module-wise Functional Contributions

This section provides additional experiments to further characterize the functional contributions of different Transformer pathways during affective adaptation. We consider two complementary settings. First, we examine how LoRA updates learned during full affective adaptation interact when they are selectively recovered and combined in the base model. Second, we retrain individual modules or module combinations under controlled settings to evaluate their independent and collaborative adaptation capacity. Together, these experiments distinguish the functional specialization observed in the final adapted model from the intrinsic adaptation capacity of individual modules.

A.6.1 Compositional Analysis of Recovered LoRA Updates

Experiment 1: Comparison among FFN pathways

We first examine whether the stronger contribution of the gating pathway observed after full affective adaptation is shared by other structurally similar FFN pathways. Specifically, we combine the recovered LoRA weights of each Attention projection with those of `gate_proj`, `up_proj`, or `down_proj`, and load each combination into the same base model.

Table 6: Compositional contributions of recovered FFN LoRA updates. Each cell reports the resulting mean performance, with the absolute change relative to the corresponding Attention-only baseline shown in parentheses.

Base Module	Base	+gate_proj	+up_proj	+down_proj
q_proj	58.72	**65.55** (+6.83)	61.27 (+2.55)	61.60 (+2.88)
k_proj	60.57	**63.98** (+3.41)	61.40 (+0.83)	60.80 (+0.23)
v_proj	57.76	**64.12** (+6.36)	58.92 (+1.16)	58.76 (+1.00)
o_proj	59.93	**62.62** (+2.69)	60.12 (+0.19)	59.31 (−0.62)

As shown in Table 6, adding the recovered `gate_proj` updates consistently produces larger performance gains across different Attention projections than adding `up_proj` or `down_proj`. This suggests that, in the fully adapted model, the updates learned by the gating pathway make a stronger compositional contribution to affective performance.

Experiment 2: Cross-structure composition with Attention pathways

We next compare combinations formed by two Attention pathways with those formed by an Attention pathway and `gate_proj`. Starting from the performance obtained by recovering a single Attention projection, we separately add another Attention projection or the recovered `gate_proj` updates.

Table 7: Comparison of recovered LoRA update compositions across Attention and gating pathways. Each cell reports the resulting mean performance, with the absolute change relative to the corresponding Attention-only baseline shown in parentheses.

Base Module	Base	+q_proj	+k_proj	+v_proj	+gate_proj
q_proj	58.72	–	60.09 (+1.37)	58.68 (−0.04)	**65.55** (+6.83)
k_proj	60.57	60.09 (−0.48)	–	58.96 (−1.61)	**63.98** (+3.41)
v_proj	57.76	58.68 (+0.92)	58.96 (+1.20)	–	**64.12** (+6.36)
o_proj	59.93	**62.63** (+2.70)	61.36 (+1.43)	60.40 (+0.47)	62.62 (+2.69)

Table 7 shows that combinations between Attention projections generally provide smaller and less consistent gains, whereas adding `gate_proj` produces more stable improvements. These results indicate that the contribution of the recovered gating updates cannot be explained simply by combining additional adapted modules.

Experiment 3: Marginal contribution under stronger Attention recovery

We further test whether this contribution persists when a larger fraction of the learned Attention updates is already recovered. We first load pairs of Attention LoRA updates into the base model and then additionally recover the LoRA weights of gate_proj, up_proj, or down_proj.

Table 8: Marginal contributions of recovered FFN LoRA updates under stronger Attention recovery. Each cell reports the resulting mean performance, with the absolute change relative to the corresponding Attention-pair baseline shown in parentheses.

Attention Pair	Base	+gate_proj	+up_proj	+down_proj
q+k	60.09	**66.47** (+6.38)	62.39 (+2.30)	63.04 (+2.95)
q+v	58.68	**65.53** (+6.85)	62.25 (+3.57)	61.30 (+2.62)
q+o	62.63	**64.21** (+1.58)	61.86 (−0.77)	61.41 (−1.22)
k+v	58.96	**65.85** (+6.89)	62.01 (+3.05)	61.91 (+2.95)
k+o	61.36	**62.32** (+0.96)	60.77 (−0.59)	61.02 (−0.34)
v+o	60.40	**62.42** (+2.02)	60.38 (−0.02)	58.11 (−2.29)

As shown in Table 8, gate_proj continues to provide a relatively stable positive gain across different Attention combinations, whereas the gains from up_proj and down_proj are less consistent. This result further supports the pronounced functional contribution of the gating pathway in the final affectively adapted model.

A.6.2 Controlled Analysis of Module Adaptation Capacity

The recovery experiments above characterize the functional organization of updates learned during full affective adaptation. However, they do not establish whether the observed differences arise from intrinsic properties of the modules or from the optimization process. We therefore conduct additional controlled training experiments.

Experiment 1: Parameter-matched single-module adaptation We first control for differences in trainable parameter count and independently adapt each projection while keeping the remaining model parameters frozen. The LoRA configurations are adjusted such that the numbers of trainable parameters are approximately matched across modules.

As shown in Table 9, FFN projections consistently outperform Attention projections under comparable parameter budgets. Although gate_proj achieves the highest mean performance, up_proj and down_proj also exhibit strong independent adaptation capability. These results suggest that the advantage of the FFN is not simply attributable to differences in trainable parameter count.

Experiment 2: Collaborative adaptation among FFN pathways We next examine whether gate_proj retains a unique advantage when different FFN pathways are jointly trained with the same Attention projection. For each Attention projection, we separately pair it with gate_proj, up_proj, or down_proj and train the corresponding module pair under the same setting.

As shown in Table 10, all three FFN projections consistently improve affective performance when jointly trained with Attention, and their overall gains are comparable. In contrast to the recovery results, gate_proj does not exhibit a stable exclusive advantage in this controlled training setting. This result suggests that the stronger gate preference observed after full adaptation is unlikely to arise solely from an intrinsic advantage of the gating pathway.

Experiment 3: Cross-structure collaborative adaptation We further compare collaboration between two Attention projections with collaboration between an Attention projection and gate_proj. Each module pair is jointly trained from the same base model under identical training conditions.

Table 11 shows that both Attention–Attention and Attention–gate combinations substantially improve affective performance. Although combinations involving gate_proj achieve a slightly higher average gain, this advantage is not consistently observed across all module pairs. These results suggest that both within-structure and cross-structure module combinations can effectively support affective adaptation.

Experiment 4: Marginal FFN contribution under enhanced Attention adaptation

Table 9: Performance comparison of single-module adaptation under approximately matched trainable-parameter budgets across all seven projection modules. *Drop* denotes the absolute decrease in mean performance relative to AffectGPT. Higher Mean and lower Drop indicate better performance preservation.

Group	Module	Emotion Recognition				Sentiment Analysis				Overall	
		MER2023	MER2024	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMS v2	Mean ↑	Drop ↓
Reference	AffectGPT	78.54	78.80	55.65	60.54	81.30	80.90	88.49	86.18	76.30	–
Attention	q_proj	55.33	53.99	49.94	42.87	76.40	63.36	81.88	81.25	63.13	13.17
	k_proj	60.35	61.39	55.87	49.01	79.55	70.69	77.35	79.27	66.68	9.62
	v_proj	52.34	56.50	45.94	46.32	81.25	71.93	78.43	76.07	63.60	12.70
	o_proj	59.23	60.84	49.65	49.80	81.85	64.66	81.43	80.49	65.99	10.31
FFN	gate_proj	71.63	76.11	56.48	58.69	80.88	78.05	84.42	83.22	**73.68**	**2.62**
	up_proj	70.78	75.30	52.85	57.55	75.51	79.09	85.76	83.44	72.53	3.77
	down_proj	75.42	74.91	54.58	57.41	75.73	78.25	87.77	83.34	73.43	2.87

Table 10: Collaborative adaptation performance with different FFN pathways. Each cell reports the resulting mean performance, with the absolute gain over the corresponding single-module baseline shown in parentheses.

Base Module	Base	+gate_proj	+up_proj	+down_proj
q_proj	68.98	72.74 (+3.76)	**74.44** (+5.46)	71.88 (+2.90)
k_proj	71.89	73.55 (+1.66)	73.74 (+1.85)	**73.77** (+1.88)
v_proj	61.98	**74.07** (+12.09)	73.78 (+11.80)	73.77 (+11.79)
o_proj	65.66	73.67 (+8.01)	**74.31** (+8.65)	71.85 (+6.19)

Finally, we examine whether FFN pathways continue to provide additional affective adaptation capacity when the Attention side has already been strengthened. We first jointly train pairs of Attention projections and then add gate_proj, up_proj, or down_proj to the training configuration.

As shown in Table 12, the marginal gains from adding FFN pathways generally decrease as the Attention configuration becomes stronger. Nevertheless, gate_proj achieves the highest average gain and remains positive across most combinations. This result suggests that the gating pathway retains relatively stable complementary capacity, although its advantage becomes less pronounced as more pathways participate in joint adaptation.

A.7 Downstream Dataset Overview

Across our experiments, we use 13 datasets spanning affective and non-affective tasks. The affective datasets include MER2023, MER2024, IEMOCAP, MELD, CMU-MOSI, CMU-MOSEI, CH-SIMS, CH-SIMS v2, Emilya, KDAE, and EGBM, covering multimodal emotion recognition, sentiment analysis, and body-motion-based emotion recognition. We additionally use GSM8K and MATH for mathematical reasoning. Brief descriptions of these datasets are provided below.

MER2023 (Lian et al., 2023) and **MER2024** (Lian et al., 2024) are widely used benchmarks for Chinese multimodal emotion recognition, where MER2024 extends MER2023. Both datasets are derived from movies and TV series and are segmented using multiple techniques to ensure that each clip contains a single speaker with relatively complete speech content. They annotate six emotion categories: worry, happiness, neutral, anger, surprise, and sadness.

IEMOCAP (Busso et al., 2008) is one of the most widely used datasets for multimodal emotion recognition. It comprises five sessions, each featuring a male and a female actor performing in a laboratory setting. The dataset includes the following emotion labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise, and other. Following previous research (Poria et al., 2017), in our experiments, we merge surprise into the happiness category.

Table 11: Cross-structure collaborative adaptation between Attention and gating pathways. Each cell reports the resulting mean performance, with the absolute gain over the corresponding single-module baseline shown in parentheses.

Base Module	Base	+q_proj	+k_proj	+v_proj	+o_proj	+gate_proj
q_proj	68.98	–	72.49 (+3.51)	**72.90** (+3.92)	72.82 (+3.84)	72.74 (+3.76)
k_proj	71.89	72.49 (+0.60)	–	**73.89** (+2.00)	72.69 (+0.80)	73.55 (+1.66)
v_proj	61.98	72.90 (+10.92)	73.89 (+11.91)	–	73.65 (+11.67)	**74.07** (+12.09)
o_proj	65.66	72.82 (+7.16)	72.69 (+7.03)	73.65 (+7.99)	–	**73.67** (+8.01)

Table 12: Marginal contributions of FFN pathways under stronger Attention adaptation. Each cell reports the resulting mean performance, with the absolute gain over the corresponding Attention-pair baseline shown in parentheses.

Attention Pair	Base	+gate_proj	+up_proj	+down_proj
q+k	72.49	**74.69** (+2.20)	73.30 (+0.81)	71.08 (−1.41)
q+v	72.90	**74.32** (+1.42)	73.05 (+0.15)	71.55 (−1.35)
q+o	72.82	**74.41** (+1.59)	73.52 (+0.70)	73.39 (+0.57)
k+v	73.89	73.30 (−0.59)	72.42 (−1.47)	**73.96** (+0.07)
k+o	72.69	73.59 (+0.90)	**74.06** (+1.37)	73.57 (+0.88)
v+o	73.65	73.81 (+0.16)	**74.54** (+0.89)	73.89 (+0.24)

MELD (Poria et al., 2019) extends the text-centric EmotionLines dataset by incorporating audio and visual modalities. The data are sourced from the TV series Friends and annotated with seven emotion categories: anger, joy, sadness, neutral, disgust, fear, and surprise.

CMU-MOSI (Zadeh et al., 2017) and **CMU-MOSEI** (Bagher Zadeh et al., 2018) consist of opinion videos collected from online platforms, with CMU-MOSEI serving as an expanded version of CMU-MOSI in both scale and topical diversity. Each sample is annotated with a sentiment intensity score in $[-3, 3]$, where -3 denotes extremely negative sentiment and $+3$ denotes extremely positive sentiment.

CH-SIMS (Yu et al., 2020) and **CH-SIMS v2** (Liu et al., 2022) focus on affective expressions grounded in Chinese cultural contexts. The data are collected from movies, TV series, and variety shows. Both datasets provide sentiment intensity annotations in $[-1, 1]$, where -1 indicates extremely negative affect and $+1$ indicates extremely positive affect.

Emilya (Fourati & Pelachaud, 2018) comprises 8,206 samples. Eleven actors performed eight emotions: Neutral, Joy, Anger, Panic, Fear, Anxiety, Sadness, and Shame, across seven daily actions, including sitting, walking, and lifting. Data were recorded using the Xsens MVN system, capturing 28 3D joints at a frequency of 120 Hz.

KDAE (Sapiński et al., 2019) was collected using a portable motion capture system that tracked 72 body markers at a frequency of 125 Hz. This dataset includes 1,402 recordings of seven emotions: Happiness, Sadness, Neutral, Anger, Disgust, Fear, and Surprise, performed by 22 semi-professional actors. For our analysis, we excluded hand joints and focused on 24 full-body joints.

EGBM (Zhang et al., 2020) contains 560 samples captured by the Kinect V2 camera, each with a frame rate of 30 Hz. The data included 16 professional actors performing seven different emotions, including Happiness, Sadness, Neutral, Anger, Disgust, Fear and Surprise. Each emotion category contains 80 samples, each consisting of 3D coordinates provided by 25 joints.

GSM8K (Cobbe et al., 2021) comprises 8.5K high-quality middle-school mathematics problems, including 7.5K training examples and 1K test examples. Each problem requires 2–8 reasoning steps and can be solved using basic arithmetic operations, including addition, subtraction, multiplication, and division.

MATH (Hendrycks et al., 2021) comprises 12,500 challenging mathematics competition problems spanning a broad range of domains, including algebra, geometry, and calculus. Each problem is accompanied by a detailed solution process.

A.8 PERFORMANCE METRICS

To evaluate different module weight loading strategies, training schemes, and the GET framework, we follow the evaluation protocol associated with each affective model. Specifically, for AffectGPT (Lian et al., 2025), we use the *hit rate* to evaluate performance on its downstream affective tasks. For Emotion-LLaMA, performance is evaluated using the F1 score, whereas for EAI-LLM, we use accuracy (ACC). For each model, we first evaluate its corresponding downstream tasks using the designated metric and then compute the arithmetic mean across these tasks as the overall performance (*Mean*). The *Mean* is computed as:

$$\text{Mean} = \frac{1}{n}\sum_{d=1}^{n} s_d, \tag{1}$$

where s_d denotes the performance score (e.g., HIT, WAF) on the d-th dataset, and *Mean* represents the arithmetic average of the performance across the n datasets.

For AffectGPT, the *hit rate* is computed as:

$$\text{hit} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[G_{w_k}(y_i) \in G_{w_k}(\hat{Y}_i)\right], \tag{2}$$

where $\mathbb{I}[\cdot]$ denotes an indicator function. Since emotion description tasks may produce free-form outputs $\hat{Y}_i$ containing multiple emotion labels, including finer-grained categories beyond the basic emotion set, the hit rate measures whether the predicted output contains the ground-truth basic emotion label y_i.