DOI: 10.1111/cgf.70368

EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

Enhancing Robust Category-Agnostic Pose Estimation through Multi-Modal Feature Alignment

Boxuan Li[1] and Juan Liu[1†]

[1]Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University

Abstract

Category-Agnostic Pose Estimation (CAPE) aims to detect keypoints for objects of any category using only a few labeled samples, making it a challenging yet crucial task for general-purpose visual understanding. Existing methods rely on either visual or textual inputs, but the lack of cross-modal interaction limits generalization. Without a unified input representation, solely using visual features hinders consistent prediction of same-type keypoints, while fixed textual representations fail to capture the diverse characteristics of same-type keypoints, leading to coarse and over-generalized outputs. To address these limitations, we propose two multi-modal frameworks that perform visual-textual integration at both the feature and decision levels. Our feature-level module leverages cross-modal attention to align and enhance keypoint representations, while the decision-level fusion adaptively combines modality-specific predictions through a modality-consistency loss. Experiments on the large-scale MP-100 dataset demonstrate that our method surpasses existing baselines in both accuracy and robustness. Under the challenging 1-shot setting, our model achieves a 0.58% improvement in PCK0.2 over the state-of-the-art CAPE method.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision; Machine learning;*** *Scene understanding; Multi-task learning; Neural networks;*

1. Introduction

Traditional pose estimation methods often rely on category-specific models that are tailored to individual object classes. While such models can achieve satisfactory performance on known categories, they often struggle to generalize to previously unseen objects due to their limited flexibility. Moreover, extending these approaches to new categories typically requires significant manual annotation of keypoints, resulting in a substantial labeling burden. To address these limitations, the task of Category-Agnostic Pose Estimation (CAPE) has been proposed in recent years. CAPE aims to eliminate the dependence on category labels by leveraging structural priors of objects, enabling accurate keypoint prediction for novel categories. This paradigm greatly enhances the scalability and practical applicability of pose estimation methods.

With the emergence of large-scale vision foundation models such as the Segment Anything Model (SAM) [ZYZ*23]and the Recognize Anything Model (RAM) [ZHM*23], CAPE can be seen as a step toward a Pose Anything Model (PAM), extending the "anything" paradigm to keypoint estimation. CAPE methods aim to match support keypoints with their corresponding locations in a query image. Prior approaches rely on single-view representations of keypoints for support-query matching either textual descriptions or local visual features.

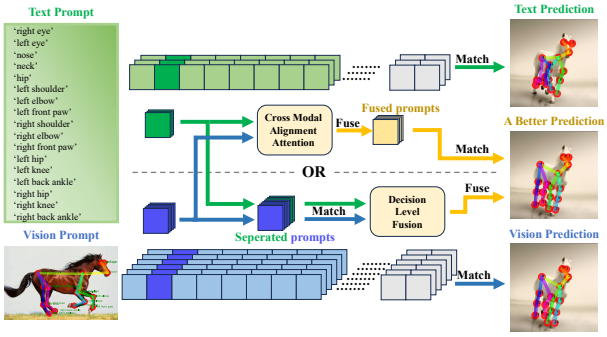

Figure 1: *Graphical Abstract: The text features and visual features in the embedding space exhibit a one-to-one correspondence. Since CAPE is fundamentally a feature matching task, utilizing a richer fused representation can result in a more robust prediction.*

Category-Agnostic Pose Estimation (CAPE) is an emerging direction in 2D pose estimation that aims to predict keypoints for ar-

† Corresponding author: Juan Liu.

bitrary object categories without category-specific training. Unlike traditional models (e.g. DeepPose [TS14], OpenPose [CHS*19], YOLO-Pose [MNMP22]), CAPE lifts category constraints, enabling keypoint localization for novel and unseen categories with only a few annotated support images. This greatly reduces the reliance on extensive manual annotations.

In recent years, several innovative frameworks have been proposed to address this challenge. Tab.1 summarizes key research progress in CAPE.

2. RELATED WORK

2.1. Traditional Pose Estimation

Pose estimation, a core task in computer vision, aims to recover keypoint locations from images. Traditional methods rely on category-specific keypoints and structural priors, often using CNNs with regression or heatmap techniques (Toshev & Szegedy, 2014; Newell et al., 2016). While effective within training categories, these approaches struggle to generalize to novel or variant categories due to fixed keypoint definitions.

As research progressed, pose estimation expanded beyond humans to objects like animals and vehicles. Customized architectures, such as DeepPose and Hourglass Network, were developed for specific categories but remain category-dependent. This limitation has driven growing interest in designing category-agnostic pose estimation models.

2.2. Category-Agnostic Pose Estimation

Category-Agnostic Pose Estimation (CAPE) is an emerging research direction within 2D pose estimation, aiming to predict keypoints for objects of arbitrary categories without requiring category-specific training. Unlike traditional pose estimation models, which typically focus on predefined object categories (e.g., humans, animals, or vehicles), CAPE seeks to remove category constraints, enabling models to localize keypoints for novel object categories with only a few support examples, and to generalize to unseen categories with minimal supervision.

2.2.1. Vision-based CAPE

Xu et al. proposed the CAPE task and introduced the Pose Matching Network (POMNet [XJZ*22]), the first CAPE framework that formulates keypoint detection as a matching problem via the Keypoint Interaction Module (KIM). They also built the MP-100 dataset and demonstrated that POMNet outperforms few-shot baselines like ProtoNet [SSZ17] and MAML [FAL17].

Shi et al. introduced CapeFormer [SHM*23], a two-stage framework addressing matching ambiguity by generating keypoint proposals and refining them with a Transformer decoder. Cape-Former achieved SOTA performance on MP-100 through query-support joint optimization and the use of support keypoint identifiers.

Hirschorn et al. proposed GraphCape [HA24b], which modeled keypoints as graphs using GCNs to incorporate structural priors. They further improved CapeFormer by replacing its MLP decoder with a graph-based FFN, boosting generalization under occlusions and symmetry. Subsequently, EdgeCape [HA24a] was introduced to improve the generalization of pose estimation for unseen categories through edge weight prediction-based skeleton refinement.

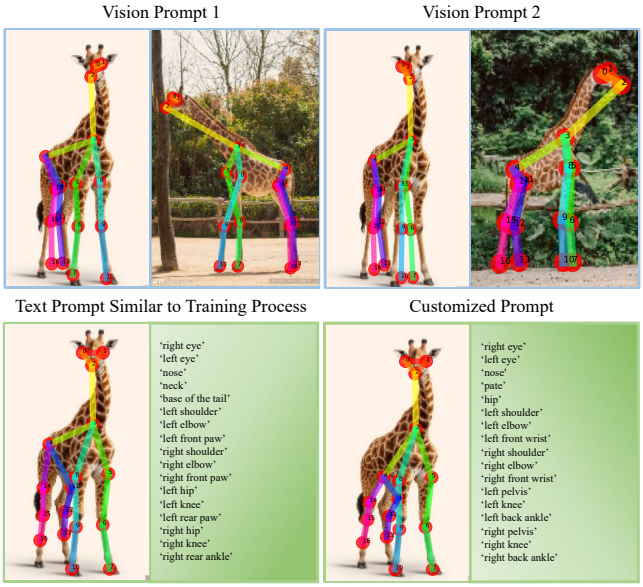

Figure 2: *Single Modal Disadvantage: The images above show the different matching results generated by different vision prompts; the images below show the consequences of not using the text descriptions generated during the training process in actual CAPEX model usage.*

2.2.2. Text-based CAPE

Rusanovsky et al. proposed CapeX [RHA25], which explored a text-based approach to category-agnostic pose estimation by leveraging textual descriptions of keypoints and modeling their relationships through graph-based structures. Their method employed open-vocabulary learning to align textual and query features, utilizing a unified representation as a proxy for semantically similar keypoints. This design effectively enhanced the correspondence between the support and query samples. However, CapeX relied on category-shared keypoint descriptions, which constrained its ability to generalize to novel categories. For example, it used identical textual labels for keypoints across all animal poses. While this approach yielded accurate results when predefined templates were used, its performance degraded significantly when user-provided descriptions were ambiguous or deviated from those seen during training, as illustrated in the bottom of Fig.2.

Kim et al. further extended text-driven CAPE by introducing large language models. Unlike CapeX, which relied on predefined keypoint descriptions, CapeLLM predicts keypoints through LLM prompting, completely eliminating the need for support images and keypoint annotations. However, the key innovation of CapeLLM is its ability to process unstructured user inputs, eliminating the need for predefined textual prompts as required by CapeX. In addition, the CapeLLM team adopted a stronger feature extraction backbone, which has led to state-of-the-art results in the CAPE field.

Table 1: *Work of current CAPE models: Overview of recent CAPE task models: modality usage and key contributions.*

Method	Vision-based	Text-based	Contribution
POMNet [XJZ*22]	✗	✗	First CAPE framework, MP-100 dataset. GitHub
CapeFormer [SHM*23]	✓	✗	Two-stage framework with similarity-based proposal and refinement. GitHub
GraphCape [HA24b]	✓	✗	Graph-based keypoint modeling. GitHub
CapeX [RHA25]	✗	✓	Using textual description features for keypoint matching. GitHub
CapeLLM [KCK24]	✗	✓	LLM-based zero-shot keypoint prediction. GitHub

Fig.2 illustrates the variation in model predictions under different uni-modal inputs. For objects of the same category, substantial differences in visual features across support images can lead to inconsistent keypoint matching results on the query image, depending on the choice of support. In contrast, using a fixed textual description provides the CAPE model with a unified representation for matching with query images, thus improving consistency. However, such a "specific vocabulary matching strategy poses practical limitations. In real-world scenarios, users may be unaware of the exact textual descriptions used during model training. When user-defined descriptions deviate from those seen during training, the model is prone to substantial keypoint prediction errors.

To address this, we propose a holistic multi-modal framework that uses a support graph where each node is enriched with open-vocabulary textual descriptions. In contrast to prior approaches that either rely solely on visual inputs or fixed textual description, our method jointly leverages the abstraction of language and the discriminative power of visual features. This cross-modal interaction enables robust matching between query keypoint appearance and support text. As shown in Graphical Abstract, our model demonstrates greater robustness in real world scenarios where descriptions may be vague or incomplete.

Our main contributions are as follows:

- **Research highlight 1:** To address the modality gap between textual descriptions and visual representations, we propose the Cross-Modal Alignment Attention (CMAA) mechanism, which performs fine-grained feature-level fusion between local visual features and textual descriptions
- **Research highlight 2:** We introduce a dual-head decision-level architecture with a visual feature and a textual feature matching Transformers, and design a modality-consistency loss to align their predictions.
- **Research highlight 3:** Experiments on the MP-100 benchmark show that our method outperforms GraphCape and CapeX in the CAPE evaluation indicators PCK0.2.
- **Research highlight 4:** When annotating unseen categories, our method fully leverages visual and textual features to enhance the robustness and accuracy of keypoint matching, thereby improving annotation quality and reducing labeling cost.

3. Method

We explore multi-modal fusion from two perspectives: (1) At the feature level, we design a novel Cross-Modal Alignment Attention(CMAA) module to bridge the semantic gap between textual description features and visual features, enabling the fused representations to be more effective for keypoint matching; (2) At the decision level, we extend the CapeX framework by introducing an additional visual feature matching path alongside the textual description matching path. A new cross-modal consistency loss is employed to align the predictions from both modalities during inference, thereby further improving the matching accuracy.

3.1. Feature Level Fusion

Visual and textual modalities provide complementary perspectives for understanding keypoint features in the support set. While images capture spatial structures and appearance details, text encodes high-level semantics and attribute descriptions. The primary challenge lies in effectively integrating these modalities to enable meaningful cross-modal reasoning. Through feature-level fusion, the model acquires enriched structural and semantic information, particularly beneficial for tasks requiring precise modal alignment such as keypoint matching. Feature-level cross-modal interaction enables dynamic learning of visual-textual associations, while attention mechanisms focus on semantically relevant text tokens to guide and refine visual representations.

To this end, we propose a feature-level fusion module based on cross-modal attention, which aligns and integrates visual features from support images with corresponding textual descriptions. This design facilitates deeper interaction across modalities and yields more discriminative joint representations for downstream matching tasks. Our feature-level fusing framework is shown in Fig.3.

3.2. Fusing Strategy

In the CAPE task, the ability to jointly model visual and textual modalities is critical for keypoint localization in unseen categories. CapeX previously attempted to align visual and textual features using CLIP [RKH*21](short for Contrastive Language–Image Pre-training) as a feature extraction backbone. However, the significant semantic gap between high-dimensional spaces of different modalities leads to instability in optimization and poor convergence, even when the feature shapes are aligned. The results in Tab.2. validate this conclusion from the perspective of quantitative metrics.

Specifically, CapeX and GraphCape extract keypoint representations via distinct backbones: a vision encoder (SwinTransformer V2 [LLC*21]) and a text encoder (GTE-base-en-v1.5 [LZZ*23]). Visual input images ([B,3,256,256]) are transformed into $[B, C_V, H, W]$, flattened, and indexed to obtain up to 100 keypoint vectors per sample ($[B,\ C_V,\ K]$). Meanwhile, textual labels (e.g., "nose", "right ankle") are embedded into $[B, C_T, K]$.

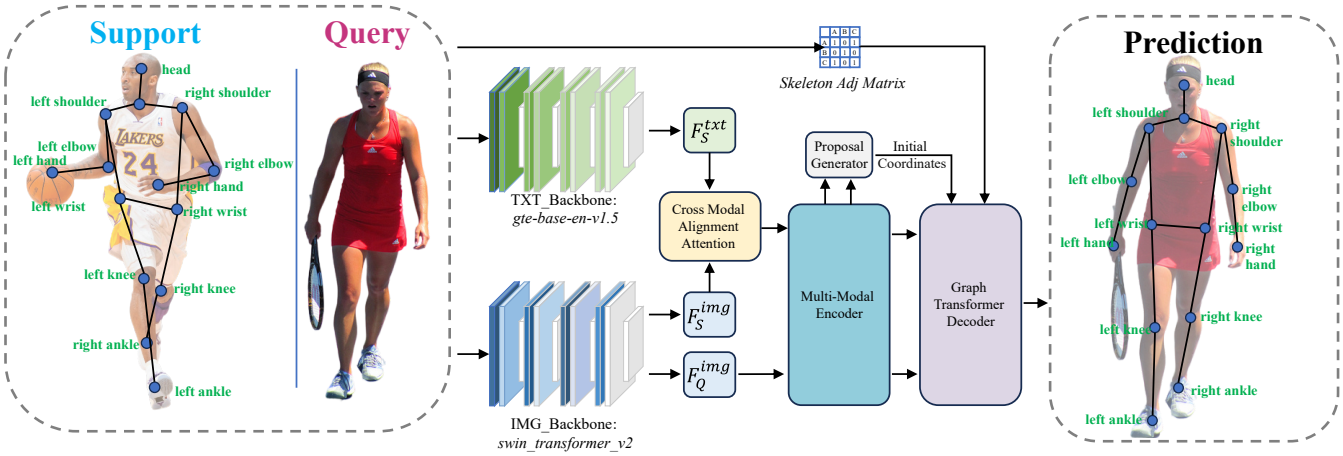

Figure 3: *Feature Level Fusing Framework: The core of feature level fusing is the Cross Model Fusion Attention module, we try to utilize two different modalities by fusing rather than using a contrastive loss to decrease the distance between two modal features, since the backbones of the two modalities encode the input features in different ways.*

Table 2: *The result of CapeX using CLIP and SwinV2-T as the backbone.*

Model	split1	split2	split3	split4	split5	Avg
CapeX-CLIP	95.17	88.88	87.72	88.54	91.65	90.15
CapeX-S	95.62	90.94	88.95	87.59	92.57	91.13

Despite operating in distinct representation spaces with separate feature extraction pipelines, both modalities encode identical keypoint sets per instance—establishing one-to-one correspondence. This alignment enables our multi-modal fusion strategy to synergize structure-aware visual features with semantically rich textual descriptors, thereby enhancing CAPE generalization.

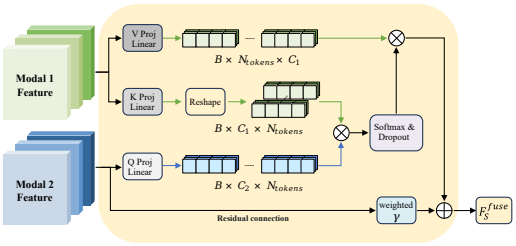

Figure 4: *Feature Level Fusing Framework: The core of feature level fusing is the Cross Model Fusion Attention module, we try to utilize two different modalities by fusing rather than using CLIP as backbone to encode image and text features into a shared embedding space, since the backbones of the two modalities encode the input features in different ways.*

3.2.1. Multi-Modal Fusing Module

To effectively learn accurate correspondences between keypoints in the support and query images, we propose a feature-level multi-modal fusion module based on cross-modal attention. This module integrates local visual features with textual descriptions, as shown in Fig.4. Instead of simply concatenating or adding modalities, our method applies token-wise attention to explicitly model the semantic alignment between visual and textual features at the feature level. Let the textual description features be denoted as $F^{\text{modal}_1} \in \mathbb{R}^{N_1 \times C}$ and the associated visual keypoint features from the support image be denoted as $F^{\text{modal}_2} \in \mathbb{R}^{N_2 \times C}$, where N is the number of keypoints and C is the feature embedding dimension. We project these features into query, key, and value spaces through learned linear transformations:

$$Q = F^{\text{modal}} W_q, K = F^{\text{modal}} W_k, V = F^{\text{modal}} W_v \tag{1}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C \times C}$ are learnable projection matrices. To identify relevant positions during token processing, the model computes query-key similarities and applies a Softmax function to obtain attention weights. Then compute cross-modal attention weights between visual and textual tokens:

$$\text{A} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \tag{2}$$

The fused feature is obtained by aggregating the attended textual information and applying residual enhancement:

$$F^{\text{fuse}}_{\text{modal}} = \text{LayerNorm}\left(F^{\text{modal}^2} + \gamma \cdot (\text{AV})\right) \tag{3}$$

where γ is a learnable scaling factor that controls the influence of the fused information on the original visual features. This cross-modal fusion module allows the model to explicitly align textual semantics with visual keypoints, improving keypoint matching accuracy. Although not strictly required during training, introducing γ enables plug-and-play usage of the module. It supports independent fine-tuning while keeping the backbone frozen, facilitating efficient transfer to new tasks.

According to the fusion formula above, textual features can be integrated into visual features, and the visual modality still plays a

dominant role. Similarly, visual features can be fused into textual features, with the textual modality taking the lead.

$$F_{S_{img}}^{fuse} = \text{LayerNorm}\left(F_S^{img} + \gamma \cdot (A \times F_S^{txt} \times W_V)\right) \tag{4}$$

3.2.2. Decision Level Supervisory Signal

Feature-level fusion enables fine-grained cross-modal interaction, but misalignment occurs under noisy or semantically divergent inputs. Thus, we adopt decision-level fusion, deferring integration until after forming modality-specific predictions.

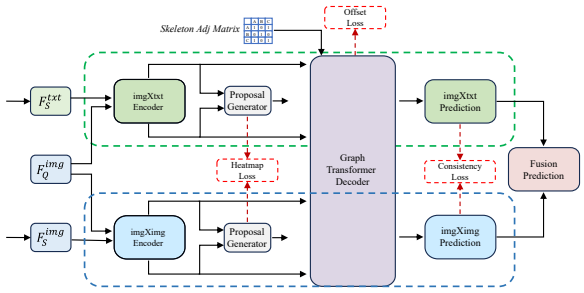

Figure 5: *Decision Level Fusing Framework: Two encoders with identical architectures process different features from distinct modalities. A consistency loss is applied to align their outputs. The two encoders("imgXimg" and "imgXtxt" encoder) are indicated by blue and green dashed boxes in the figure.*

Fig.5 shows the decision-level fusion framework. Visual and textual inputs are separately processed by "imgXimg" and "imgXtxt" encoders to generate modality-specific region proposals. These are then fused using a unified graph-based decoder. After a pretrained backbone (like Swin Transformer) extracts features from the images, the Transformer Encoder applies self-attention to model global relationships between different parts of the image. This helps the model better understand the structure and context of the object.

We introduce a consistency loss that aligns predictions across modalities, encouraging modality-invariant representations. This late-stage fusion preserves independent pipelines, reduces early interference, and enhances robustness by leveraging high-level semantic cues—ultimately improving reliability and generalization in multi-modal matching. To quantitatively enforce prediction alignment across modalities, we define the consistency loss as follows:

$$L_{\text{consistency}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{P}_{\text{text}}^{(i)} - \hat{P}_{\text{img}}^{(i)}\right\| \tag{5}$$

Here, P_{text} and P_{img} denote the predicted keypoints from the image-text and image-image pathways, respectively. The consistency loss is defined as the mean squared error (MSE) between the two sets of predictions, which is shown above. N denotes the number of valid keypoints as determined by a visibility mask. We apply this loss only to the keypoints marked as valid in the image-text modality to avoid noisy supervision.

3.3. Training Scheme

We incorporate three supervisory signals: heatmap loss, offset loss, and consistency loss. The heatmap loss guides the proposal generator by shaping similarity maps and promoting meaningful representations, while the offset loss supervises localization.

$$L_{\text{heatmap}} = \frac{1}{K \cdot H \cdot W}\sum_{i=1}^{K}\|\sigma(M_i) - H_i\| \tag{6}$$

$$L_{\text{offset}} = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{K}\left|P_i^l - \hat{P}_i\right| \tag{7}$$

where M_i is the output similarity heatmap from the proposal generator, σ is the sigmoid function, H_i is the ground-truth heatmap, P_i is the output location, and $\hat{P}_i$ is the ground-truth location.

We propose enhancements at both the feature level and the decision level. Specifi-cally, the decision-level fusion strategy incorporates a consistency loss, designed to im-prove cross-modal alignment, on top of the original objective function. The overall loss functions L_{CMA} for the feature-level model and L_{DLF} for the decision-level fusion model are defined as follows:

$$L_{\text{CMA}} = \lambda_{\text{heatmap}} \cdot L_{\text{heatmap}} + L_{\text{offset}} \tag{8}$$

$$L_{\text{DLF}} = \lambda_{\text{heatmap}} \cdot L_{\text{heatmap}} + \lambda_{\text{consistency}} \cdot L_{\text{consistency}} + L_{\text{offset}} \tag{9}$$

4. Experiment

Following prior CAPE studies, we use the MP-100 dataset for evaluation. MP-100 consists of samples collected from existing category-specific pose estimation datasets: COCO [LMB*14], 300W [SAT*16], AFLW [KWRB11], OneHand10K [WPL19], DeepFashion2 [GZW*19], AP-10K [YXZ*21], MacaquePose [LMN*21], Vinegar Fly and Desert Locust [GCN*19], CUB-200 [WBW*11], CarFusion [RVN18], AnimalWeb [KMK*20], APT-36K [YYX*22]. It contains over 18K images across 100 distinct subcategories grouped into 8 super categories: human hand/face/body, animal face/body, cloth, furniture, and vehicles with annotated keypoints. The dataset is specifically partitioned for CAPE tasks: the train/val/test splits comprise 70/10/20 categories, respectively, with approximately 200 images per category.

To evaluate the model's generalization to unseen object categories, the MP-100 dataset adopt a leave-one-split-out protocol. It comprises 100 categories, each annotated with consistent semantic keypoints. For each of the 5 splits, the dataset is partitioned into a train set (70 categories), a validation set (10 categories), and a test set (20 categories), all drawn from disjoint subsets of MP-100. This results in 5 distinct experimental configurations, ensuring that every category is evaluated exactly once as an unseen (novel) class. Such a protocol enables a comprehensive assessment of the model's ability to transfer structural knowledge learned from base categories to novel object categories during testing.

We leveraged the updated annotation files from CapeX, which standardized textual descriptions for all keypoints across all categories based on the skeleton-structured annotations originally proposed by GraphCape. These annotations provide new supervisory signals for the revised version of the MP-100 dataset.

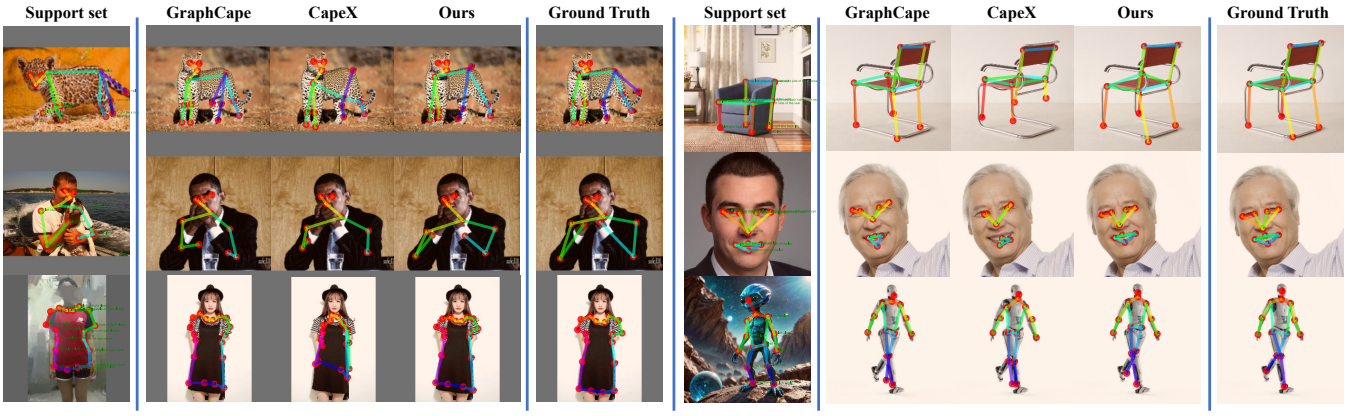

Figure 6: *Qualitative Result: We visualize keypoint predictions under the 1-shot setting. The left column shows support images with their corresponding skeletons and text descriptions. The following columns show results from GraphCape, CapeX, and our method. The last column is the ground truth.*

Our architecture is implemented in the MMPose framework [Con20]. For the modality backbones, we use the same text encoder as CapeX: `gte-base-en-v1.5`, and the same visual encoder as GraphCape: `Swin-Transformer-Tiny`. This ensures that improvements are due solely to our proposed multi-modal fusion strategy. In contrast, CapeLLM uses a stronger backbone (`DINO-v2`) [CTM*21] and focuses on leveraging LLMs to solve the CAPE task. Our work instead highlights the effectiveness of a multi-modal fusion mechanism for CAPE prediction.

Tab.3 presents the key hyperparameters table.

Table 3: *Key model hyperparameters(grouped by stage).*

Stage	Hyperparameter	Value
Training	total_epochs	200
	learning_rate	1e-5
	optimizer_type	Adam
Model Architecture	visual_backbone	Swin-T V2
	textual_backbone	gte-base-en-v1.5
	img_size	[256, 256]
	embed_dim	96
	depths	[2, 2, 18, 2]
	num_heads	[3, 6, 12, 24]
	drop_path_rate	0.3
Transformer Head	d_model	256
	nhead	8
	encoder/decoder_layers	3/3
	dim_feedforward	768
	dropout	0.1
	activation	relu
Data Processing	batch_size	8
	num_shots	1

4.1. Benchmark Results

We compared GraphCape, CapeX, feature-level fusion methods, and decision-level fusion methods under the 1-shot setting. In our framework, supporting information is provided by a single sample of textual descriptions and visual features. We show a quantitative comparison between our method and previous CAPE methods: GraphCape, CapeX in Fig.6 The results on the MP-100 dataset under 1-shot are reported in Tab.4.

Table 4: *PCK performance under 1-shot setting. "-S" means Swin Transformer V2 backbone.*

Model	split1	split2	split3	split4	split5	Avg
ProtoNet	46.05	40.84	49.13	43.34	44.54	44.78
MAML	68.14	54.72	64.19	63.24	57.20	61.50
POMNet	84.23	78.25	78.17	78.68	79.17	79.70
CapeFormer-S	92.88	89.11	89.16	87.19	88.73	89.41
GraphCape-S	94.73	89.79	90.69	88.09	90.11	90.68
CapeX-S	95.62	90.94	88.95	89.43	92.57	91.50
Ours-S	**95.95**	**91.24**	**91.00**	**89.61**	**92.60**	**92.08**

These results demonstrate the advantages of multi-modal fusion, showing that our approach consistently achieves superior keypoint localization accuracy compared to prior CAPE methods. Fig.7 shows Quantitative prompt-robustness evaluation, We use different textual descriptions to compare the results with CapeX as the baseline. While the predicted keypoints vary with the prompts, our model produces more acceptable and consistent predictions under new prompts compared to CapeX.

4.2. Computational Efficiency and Model Complexity

To evaluate the efficiency and scalability of our method, we compare the computational cost, model size, and inference speed with existing baselines including CapeX and GraphCape. Tab.5 summarizes the number of FLOPs, total parameters, average inference time, and frames per second (FPS) across different models.

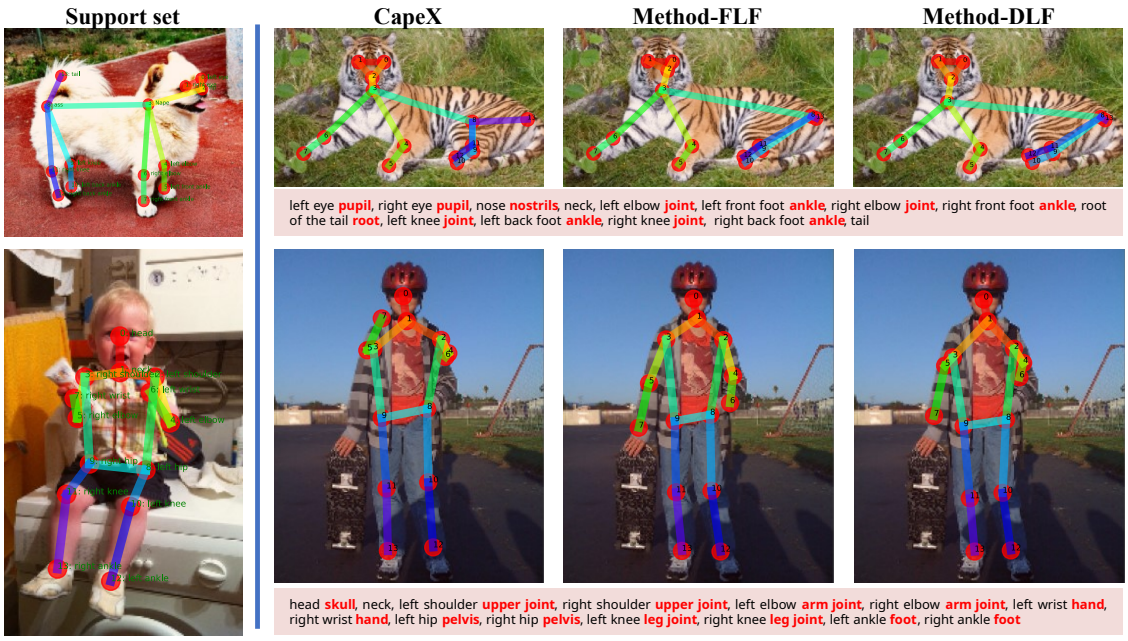

Figure 7: *Quantitative prompt-robustness evaluation: We use different textual descriptions to perform prediction and compare the results with CapeX as the baseline. While the predicted keypoints vary with the prompts, our model produces more acceptable and consistent predictions under new prompts compared to CapeX.*

Our approach introduces two variants: Method-FLF(Feature-level fusion) and Method-DLF(Decision-level fusion). Despite integrating additional modules—particularly the cross-modal attention mechanism—our models incur only a modest increase in computational cost. Specifically, the additional component contributes approximately 8.64M FLOPs and introduces only a single learnable scalar parameter, resulting in a negligible increase in model size. Compared to CapeX, Method-FLF shows a moderate increase in FLOPs to 24.456 GFLOPs, while maintaining nearly the same parameter count (195.715M). Similarly, Method-DLF has 24.827 GFLOPs and 203.742M parameters, reflecting a slightly larger model due to the decision-level fusion strategy.

Table 5: *Comparison of computational cost, model size, and inference speed. Method-FLF stands for Feature-level fusion and Method-DLF stands for Decision-level fusion.*

Model	FLOPs (G)	Params (M)	Time (ms)	FPS
GraphCape	17.657	58.742	61.85	16.15
CapeX	15.813	195.519	**38.27**	**26.13**
Method-FLF	24.456	195.715	61.35	16.30
Method-DLF	24.827	203.742	68.85	14.52

In terms of runtime, CapeX achieves the fastest inference (38.27 ms, 26.13 FPS), but our models still run in real-time: Model-1 at 61.35 ms (16.30 FPS) and Model-2 at 68.85 ms (14.52 FPS), comparable to GraphCape (61.85 ms). Despite higher FLOPs, our multi-modal fusion introduces minimal overhead and offers a strong efficiency-accuracy trade-off. FPS values are averaged over 100 runs on an RTX 3060.

4.3. Ablation Experiment

Since our proposed model relies on both visual and textual modalities, the key difference between the 2 fusion strategies lies in how the modalities are integrated—either at the feature level via early fusion (FLF) or at the decision level via a consistency loss that aligns predictions from unimodal branches. Therefore the CMAA cannot be evaluated in isolation. Thus, our ablation compares four settings: unimodal visual, unimodal textual, multi-modal FLF, and multi-modal DLF.

- **Unimodal:**
  - **Visual-only:** Uses only visual features to match.
  - **Textual-only:** Uses only textual features to match.
- **Multi-modal (Visual + Textual):**
  - **Feature-level Fusion:** Method-FLF that integrates visual and textual features early via the proposed CMAA module.
  - **Decision-level Fusion:** Method-DLF's fusion occurs at the decision level, guided by a cross-modal consistency loss.

Table 6: *Ablation Results: Comparison of results generated from different modality inputs. Method-FLF/DLF denote feature-level and decision-level fusion.*

Model	split1	split2	split3	split4	split5	Avg
Visual-only	93.82	88.97	90.32	83.37	89.93	89.28
Textual-only	94.70	89.76	89.18	87.36	91.92	90.58
Method-FLF	94.36	**90.64**	**90.39**	87.16	**92.27**	**90.96**
Method-DLF	**95.19**	89.98	89.17	**87.57**	91.76	90.93

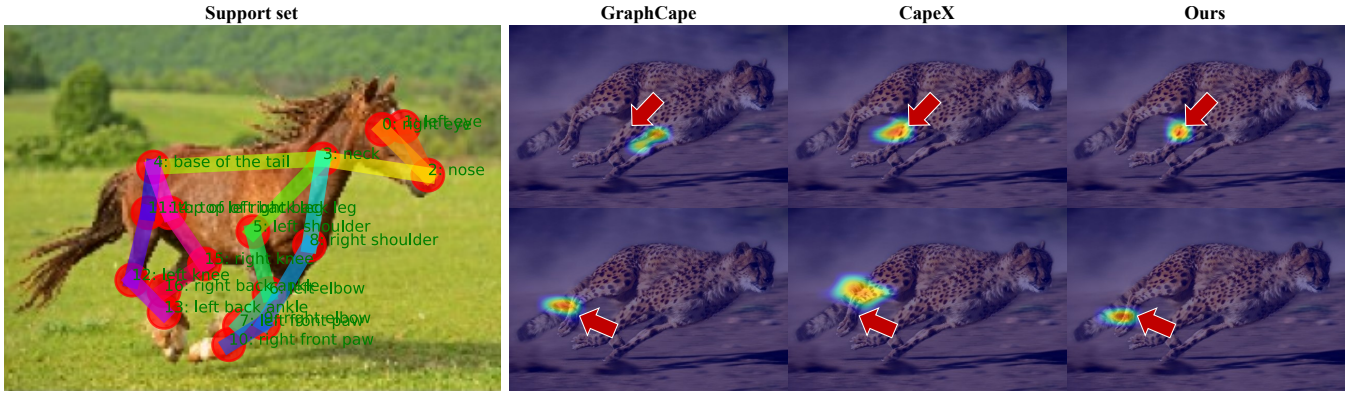

Figure 8: *Comparison of Cross-attention maps: From left to right is GraphCape, CapeX, and our model. Our model demonstrates better performance and surpasses GraphCape and CapeX by combining user-annotated keypoints with textual descriptions.*

All experiments were conducted under the same train setup without fine-tuning, using default hyperparameters and 50% of the MP-100 train set, which reduced computational and time costs.

The results demonstrate that incorporating support information—either at the feature or decision level—consistently improves model performance over unimodal baselines. Feature-level fusion excels in certain splits by integrating fine-grained visual-textual cues into intermediate representations, while decision-level fusion performs better in others due to its ability to enforce global consistency at the prediction stage. Despite variations across splits, both strategies reliably enhance keypoint localization accuracy, indicating that support-based fusion is a robust and effective mechanism.

4.4. Performance at COCO human keypoint dataset

We validated our model on the single-person subset of the COCO human keypoint detection dataset. In the human body category, our method consistently outperforms baselines across all five splits of the MP-100 dataset. As shown in Tab.7. compared to GraphCape and CapeX, we achieve higher accuracy on all splits.

Our methods consistently outperformed existing approaches, GraphCape and CapeX, across all splits (split1 to split5). Notably, Multi-modal-FLF showed significant improvements on split3 and split4, achieving 74.80% and 88.57% respectively, compared to 69.09% and 83.10 % from CapeX. On average, Multi-modal-FLF and Method-DLF achieved scores of 85.13% and 83.15%, surpassing GraphCape (71.61%) and CapeX (82.24%). These results demonstrated the superior generalization and robustness of our approach in diverse human pose estimation scenarios.

Table 7: *Comparison of different model's human body pose estimation results.*

Model	split1	split2	split3	split4	split5	Avg
GraphCape	73.36	73.50	65.03	73.79	72.35	71.61
CapeX	83.65	86.51	69.09	86.13	86.30	82.24
Method-FLF	**86.83**	**88.56**	**74.80**	**86.87**	**88.57**	**85.13**
Method-DLF	84.78	86.91	70.83	86.47	86.78	83.15

4.5. Comparison of Cross-attetion maps

Fig. 8 presents a comparison of cross-attention maps generated by GraphCape, CapeX, and our method. GraphCape exhibits misaligned and noisy attention, often focusing on irrelevant regions due to its reliance on visual support with potentially inaccurate keypoint annotations. CapeX, using only textual prompts, produces diffuse and ambiguous attention maps, reflecting its limited ability to capture fine-grained visual-textual correspondence. In contrast, our method integrates both visual and textual modalities, resulting in more focused and semantically consistent attention maps.

5. Conclusion

Our method integrates visual and textual information at feature or decision levels to fully leverage their complementary strengths. Feature-level fusion jointly encodes image and text description to improve keypoint matching across diverse categories, while decision-level fusion introduces a cross-modal consistency loss to enhance robustness and alignment. Unlike single-modality methods like CapeX and GraphCape, our approach requires both inputs during inference, demonstrating the effectiveness of multi-modal fusion without relying on backbone changes. Though limited in language-only settings, it offers a practical and efficient solution for accurate keypoint annotation, especially under low-annotation conditions. Nevertheless, requiring both visual and textual inputs during inference introduces inconvenience in real-world deployment.

Acknowledgements

This work was supported in part by the Industry-University Cooperation Project of Fujian Province under Grant 2024H6002, and in part by the Natural Science Foundation of Xiamen, China under Grant 3502Z202373023.

References

[CHS*19] CAO Z., HIDALGO MARTINEZ G., SIMON T., WEI S., SHEIKH Y. A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019). 2

[Con20] CONTRIBUTORS M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 6

[CTM*21] CARON M., TOUVRON H., MISRA I., JÉGOU H., MAIRAL J., BOJANOWSKI P., JOULIN A.: Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)* (2021). 6

[FAL17] FINN C., ABBEEL P., LEVINE S.: Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* (2017), vol. 70, pp. 1126–1135. 2

[GCN*19] GRAVING J. M., CHAE D., NAIK H., LI L., KOGER B., COSTELLOE B. R., COUZIN I. D.: Fast and robust animal pose estimation. *bioRxiv* (2019), 620245. 5

[GZW*19] GE Y., ZHANG R., WANG X., TANG X., LUO P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5332–5340. doi:10.1109/CVPR.2019.00548. 5

[HA24a] HIRSCHORN O., AVIDAN S.: Edge weight prediction for category-agnostic pose estimation, 2024. URL: https://arxiv.org/abs/2411.16665, arXiv:2411.16665. 2

[HA24b] HIRSCHORN O., AVIDAN S.: A graph-based approach for category-agnostic pose estimation, 2024. URL: https://github.com/orhir/PoseAnything, arXiv:2311.17891. 2, 3

[KCK24] KIM J., CHUNG H., KIM B.-H.: Capellm: Support-free category-agnostic pose estimation with multimodal large language models. *arXiv preprint arXiv:2411.06869* (2024). URL: https://github.com/Junhojuno/CapeLLM. 3

[KMK*20] KHAN M. H., MCDONAGH J., KHAN S., SHAHABUDDIN M., ARORA A., KHAN F. S., SHAO L., TZIMIROPOULOS G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 6937–6946. doi:10.1109/CVPR42600.2020.00697. 5

[KWRB11] KÖSTINGER M., WOHLHART P., ROTH P. M., BISCHOF H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (2011), pp. 2144–2151. doi:10.1109/ICCVW.2011.6130513. 5

[LLC*21] LIU Z., LIN Y., CAO Y., HU H., WEI Y., ZHANG Z., LIN S., GUO B.: Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 9992–10002. doi:10.1109/ICCV48922.2021.00986. 3

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Springer International Publishing, pp. 740–755. 5

[LMN*21] LABUGUEN R., MATSUMOTO J., NEGRETE S. B., NISHIMARU H., NISHIJO H., TAKADA M., GO Y., INOUE K., SHIBATA T.: MacaquePose: A Novel "In the Wild" Macaque Monkey Pose Dataset for Markerless Motion Capture. *Frontiers in Behavioral Neuroscience 14* (2021), 581154. URL: https://doi.org/10.3389/fnbeh.2020.581154, doi:10.3389/fnbeh.2020.581154. 5

[LZZ*23] LI Z., ZHANG X., ZHANG Y., LONG D., XIE P., ZHANG M.: Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281* (2023). 3

[MNMP22] MAJI D., NAGORI S., MATHEW M., PODDAR D.: Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2022), pp. 2636–2645. doi:10.1109/CVPRW56347.2022.00297. 2

[RHA25] RUSANOVSKY M., HIRSCHORN O., AVIDAN S.: Capex: Category-agnostic pose estimation from textual point explanation. In *The Thirteenth International Conference on Learning Representations* (2025). URL: https://github.com/matanr/capex, doi:https://openreview.net/forum?id=scKAXgonmq. 2, 3

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J.: Learning transferable visual models from natural language supervision. 3

[RVN18] REDDY N. D., VO M., NARASIMHAN S. G.: Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 1906–1915. doi:10.1109/CVPR.2018.00204. 5

[SAT*16] SAGONAS C., ANTONAKOS E., TZIMIROPOULOS G., ZAFEIRIOU S., PANTIC M.: 300 faces in-the-wild challenge: database and results. *Image and Vision Computing 47* (2016), 3–18. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge. URL: https://www.sciencedirect.com/science/article/pii/S0262885616000147, doi:https://doi.org/10.1016/j.imavis.2016.01.002. 5

[SHM*23] SHI M., HUANG Z., MA X., HU X., CAO Z.: Matching is not enough: A two-stage framework for category-agnostic pose estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 7308–7317. URL: https://github.com/flyinglynx/CapeFormer, doi:10.1109/CVPR52729.2023.00706. 2, 3

[SSZ17] SNELL J., SWERSKY K., ZEMEL R.: Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems* (2017), vol. 30, pp. 4077–4087. 2

[TS14] TOSHEV A., SZEGEDY C.: Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1653–1660. doi:10.1109/CVPR.2014.214. 2

[WBW*11] WAH C., BRANSON S., WELINDER P., PERONA P., BELONGIE S.: *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011. 5

[WPL19] WANG Y., PENG C., LIU Y.: Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology 29*, 11 (2019), 3258–3268. doi:10.1109/TCSVT.2018.2879980. 5

[XJZ*22] XU L., JIN S., ZENG W., LIU W., QIAN C., OUYANG W., LUO P., WANG X.: Pose for Everything: Towards Category-Agnostic Pose Estimation. *Computer Vision – ECCV 2022 13671* (2022), 398–416. Code available at https://github.com/luminxu/Pose-for-Everything. URL: https://github.com/luminxu/Pose-for-Everything. 2, 3

[YXZ*21] YU H., XU Y., ZHANG J., ZHAO W., GUAN Z., TAO D.: Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021). 5

[YYX*22] YANG Y., YANG J., XU Y., ZHANG J., LAN L., TAO D.: Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems 35* (2022), 17301–17313. 5

[ZHM*23] ZHANG Y., HUANG X., MA J., LI Z., LUO Z., XIE Y., QIN Y., LUO T., LI Y., LIU S., ET AL.: Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514* (2023). 1

[ZYZ*23] ZOU X., YANG J., ZHANG H., LI F., LI L., WANG J., WANG L., GAO J., LEE Y. J.: Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)* (2023). 37th Conference on Neural Information Processing Systems (NeurIPS), New Orleans, LA, Dec 10–16, 2023. 1