

# DESCRIBE-TO-SCORE: TEXT-GUIDED EFFICIENT IMAGE COMPLEXITY ASSESSMENT

Shipeng Liu, Zhonglin Zhang, Dengfeng Chen, Liang Zhao\*

Xi'an University of Architecture and Technology

{lsp, zhang.zhonglin215, chdengf, zhaoliang}@xauat.edu.cn

## ABSTRACT

Accurately assessing image complexity (IC) is critical for computer vision, yet most existing methods rely solely on visual features and often neglect high-level semantic information, limiting their accuracy and generalization. We introduce vision-text fusion for IC modeling. This approach integrates visual and textual semantic features, increasing representational diversity. It also reduces the complexity of the hypothesis space, which enhances both accuracy and generalization in complexity assessment. We propose the D2S (**D**escribe-to-**S**core) framework, which generates image captions with a pre-trained vision-language model. We propose the feature alignment and entropy distribution alignment mechanisms, D2S guides semantic information to inform complexity assessment while bridging the gap between vision and text modalities. D2S utilizes multi-modal information during training but requires only the vision branch during inference, thereby avoiding multi-modal computational overhead and enabling efficient assessment. Experimental results demonstrate that D2S outperforms existing methods on the IC9600 dataset and maintains competitiveness on no-reference image quality assessment (NR-IQA) benchmark, validating the effectiveness and efficiency of multi-modal fusion in complexity-related tasks. Code is available at: <https://github.com/xauat-liushipeng/D2S>

## 1 INTRODUCTION

Image complexity (Forsythe, 2009) (IC) is a fundamental factor in human visual perception, influencing aesthetic judgment, memorability (Singh & Shukla, 2017). In computer vision, accurate image complexity assessment (ICA) facilitates tasks such as automatic annotation, active learning, and hard example mining by identifying informative samples and improving learning efficiency and generalization (Feng et al., 2022). Early approaches relied on statistical features, such as fractal dimension, entropy (Li et al., 2021), and edge density (Dai et al., 2022), to capture structural richness. However, their applicability is limited by subjectivity, inconsistent standards, and poor cross-domain generalization. Recent deep learning methods, including convolutional neural networks (CNNs) and vision transformers (ViTs) (Liu et al., 2025a), leverage hierarchical feature representations to significantly improve the accuracy of complexity prediction. However, in the context of complexity modeling, these approaches still tend to rely primarily on low-level visual patterns (e.g., texture and color), while lacking explicit modeling of high-level semantics such as object count, category, or spatial relations. In addition, most of them often suffers from limited interpretability (Chen et al., 2015; Shen et al., 2024).

Recent approaches have introduced high-level semantic information (e.g., object counts (Shen et al., 2024), motion trends (Li et al., 2025)) into image complexity modeling and achieved promising results. Meanwhile, we observe that when assessing image complexity, humans attend not only to local textures and colors but also to high-level semantics, including the number, categories, and spatial relationships of objects in a scene, as well as potential events. This motivates us to raise a central question:

*How can image complexity be computationally evaluated in a human-like way, combining low-level visual features with high-level semantic information?*

---

\*Corresponding author

Therefore, we first verify whether there is any available IC information in the image caption. we attempted to use visual-language models (VLMs) (Li et al., 2022) to generate captions for images and only use the captions to learning IC. The result is incredibly surprising. This is the process shown on the right part in Figure 1, and the resulting Pearson correlation coefficient (PCC) is 0.8251 (More results are in Table 7). Based on the above issues and observation results, we continued to integrate the left and right parts in Figure 1. Attempting to use the text to guide (Text Guidance) the visual branches, we proposed the D2S (Describe-to-Score) framework.

The core of D2S is *Describe first; then Score*.

**(1) Describing.** It employs VLMs to generate image captions that capture object concepts, relationships, and structural information. **(2) Scoring.** Through vision-text alignment, caption text guides the modeling of high-level semantic information while preserving sensitivity to low-level visual features, enabling accurately text-guided ICA. The result is that text-guided PCC on the IC9600 dataset (Feng et al., 2022) has increased by 0.0146 over the pure visual method (the middle part in Figure 1).

We theoretically and empirically verify the feasibility of visual-text fusion (Park & Kim, 2023) for ICA. From an information-theoretic perspective, entropy analysis (Yang & Nataliani, 2017) shows that fused visual-text features have higher entropy than visual-only features, better approximating real IC (Section 2.2). From a generalization-theoretic perspective, Empirical Rademacher Complexity (Mohri & Rostamizadeh, 2008) reveals that semantic inputs compress visual features and reduce effective feature dimensionality, thereby tightening the error bound and improving generalization (Section 2.3). Guided by these insights, we design entropy distribution alignment (EAL) and feature alignment (FAL) mechanisms, whose effectiveness is validated through extensive experiments. D2S achieves state-of-the-art (SOTA) performance across all complexity assessment metrics (Figure 2) with significantly lower inference latency on IC9600 (Figure 3). Moreover, when transferred to no-reference image quality assessment (NR-IQA) (Mittal et al., 2012), D2S delivers competitive results, confirming its robustness across tasks. D2S has achieved the current best results on KADID-10k (Lin et al., 2019). Our main contributions are as follows:

- We propose D2S, a vision-text fusion framework that theoretically shows semantic information enriches visual representations and reduces hypothesis space complexity, improving accuracy and generalization.
- We develop entropy distribution alignment and feature alignment mechanisms to bridge the modal gap of visual-text in ICA, improving cross-modal consistency and robustness.
- Extensive experiments on IC9600, KADID-10K, and related benchmarks establish state-of-the-art performance with significantly faster inference, validating both efficiency and cross-task adaptability.

## 2 PRELIMINARY

### 2.1 TASK DEFINITION

Given an input image  $I$  from the dataset  $\mathcal{D}$ , its ground-truth complexity label is denoted as  $y \in (0, 1)$ . An arbitrary caption generation model  $g_\phi$  produces a text description  $S = g_\phi(I)$ . We then train a VLM  $f_\theta$ , which consists of a visual encoder and a text encoder, to predict IC score  $\hat{y} = f_\theta(I, S)$ . The learning objective is to minimize the error between the predicted complexity and the ground-truth:

$$\min_{\theta} \mathbb{E}_{(I, y) \sim \mathcal{D}} [\ell(f_\theta(I, S), y)] \quad (1)$$

where,  $\ell$  denotes the loss function (e.g., MSE), and  $\theta$  represents the model parameters. For the single-modal case using only image input, the prediction result is written as  $\hat{y}_v = f_\theta(I)$ . Our

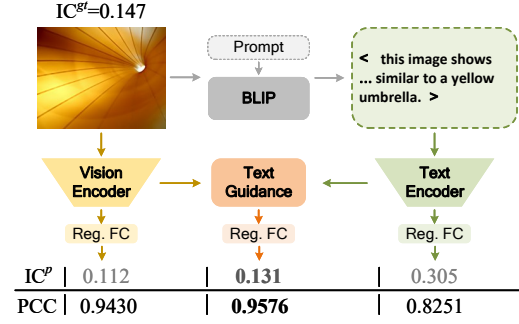


Figure 1: Text-Guided image complexity assessment.  $IC^{gt}$  and  $IC^p$  are ground-truth and predicted IC score, respectively.

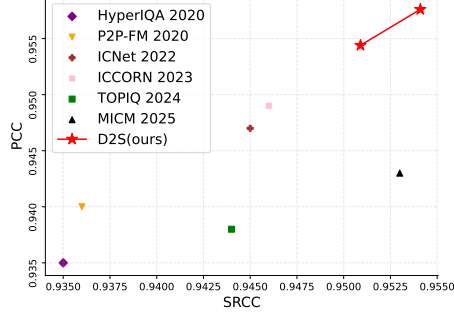


Figure 2: Accuracy comparison with SOTA (higher is better; upper right is preferred).

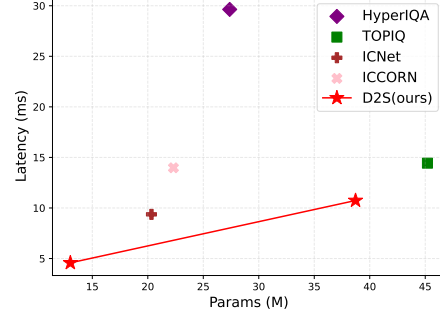


Figure 3: Efficiency comparison on RTX 3090. (lower is better; lower left is preferred).

core assumption is that multi-modal fusion performs no worse than the single-modal baseline,  $\hat{y} \geq \hat{y}_v$ . In the following, we validate this hypothesis from the perspectives of information theory and generalization theory.

## 2.2 ENTROPY AND COMPLEXITY

**Proposition 1.** *Image complexity increases with the growth of visual diversity and semantic diversity. Let the original entropy be  $H(I)$ , the entropy of visual features be  $H_v^F(I)$  and the entropy of semantic features be  $H_s^F(S)$ . The entropy after fusion satisfies:*

$$H^F(I) = \alpha H_v^F(I) + \beta H_s^F(S) > H_v^F(I) \quad (2)$$

where,  $\alpha, \beta > 0$  are weighting coefficients. **Proof** is provided in Appendix A.1.

**Implications of Proposition 1.** Semantic information complements visual information and enriches representation diversity by increasing entropy, making the fused features closer to the complexity of real images. Consequently, in complexity assessment tasks, multi-modal models are expected to outperform their single-modal counterparts.

## 2.3 GENERALIZATION VIA RADEMACHER COMPLEXITY

**Definition 1 (Empirical Rademacher Complexity (Mohri & Rostamizadeh, 2008)).** *Suppose the hypothesis space is  $\mathcal{F}$  with sample size  $n$ . If the feature representation  $\phi(I, S)$  satisfies the boundedness condition  $|\phi(I, S)| \leq B$ , and its feature dimension is  $d$ , then the empirical Rademacher complexity satisfies:*

$$\widehat{\mathcal{R}}_S(\mathcal{F}) \leq \frac{B\sqrt{d}}{\sqrt{n}} \quad (3)$$

**Lemma 1 (Dimensionality Reduction via Semantics).** *Let the visual and semantic features be denoted as  $X_v$  and  $X_s$ , respectively. If semantic features  $X_s$  exert a compressing or regularizing effect on visual features  $X_v$ , the effective dimension  $d$  will be reduced to  $d'$ ,  $d' < d$ . In this case, we have:*

$$\widehat{\mathcal{R}}'_S(\mathcal{F}) \leq \frac{B\sqrt{d'}}{\sqrt{n}} \leq \frac{B\sqrt{d}}{\sqrt{n}} \quad (4)$$

**Theorem 2 (Generalization Enhancement).** *Under the framework of visual-text fusion, if semantic inputs can reduce the effective feature dimension, then the empirical Rademacher complexity decreases, thereby enhancing the generalization capability of the model.*

**Proof.** From Lemma 1, the effective dimension is reduced from  $d$  to  $d'$ . Since the upper bound of  $\widehat{\mathcal{R}}'_S(\mathcal{F})$  is proportional to  $\frac{B\sqrt{d}}{\sqrt{n}}$ , a smaller  $d'$  leads to a tighter bound. According to statistical learning theory, a lower Rademacher complexity implies a smaller generalization error bound. Therefore, visual-text fusion is capable of improving model generalization.

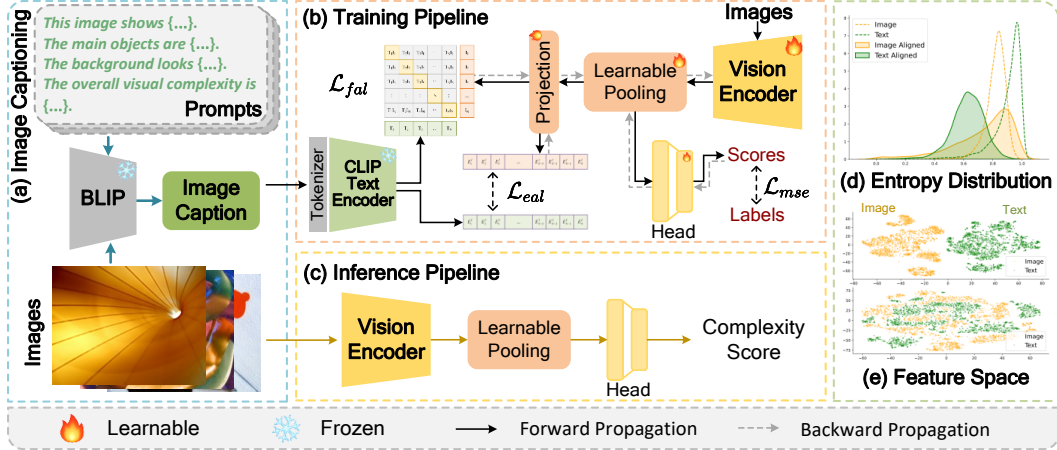


Figure 4: **Overall Architecture of the D2S framework.** (a) BLIP with a fill-in prompt generates captions. (b) Vision and text encoders extract features for regression loss  $\mathcal{L}_{mse}$  and the alignment losses  $\mathcal{L}_{fal}$ ,  $\mathcal{L}_{eal}$ . (c) At inference, only image input is used for score prediction. (d) Entropy distribution before and after alignment (Dotted line: before alignment; solid line: after alignment). (e) Feature space before (top) and after (bottom) feature alignment.

**Implications of Proposition 1 and Theorem 2.** The advantages of visual-text fusion can be summarized in two aspects. (1) Increasing representational diversity, thereby better approximating the true complexity of images. (2) Reducing the effective hypothesis space dimension, thereby enhancing model generalization. This theoretical analysis can fully demonstrate that the use of multi-modal methods in complexity assessment and related tasks will perform better than the single-modal benchmark methods.

### 3 METHOD

#### 3.1 DESCRIBE-TO-SCORE

Our proposed D2S framework aims to jointly model low-level visual features and high-level semantic information through text guidance, thereby achieving robust ICA. Figure 4 is the overall workflow. Given an input image, a pre-trained BLIP first generates captions. The vision encoder then extracts visual features, while the text encoder extracts textual features. The vision-text alignment module of D2S consists of entropy distribution alignment and feature alignment. In this process, textual information acts as a semantic teacher, guiding the visual encoder to learn semantic-aware features, but **it is not directly used in the computation of the final complexity score**. The complexity score is produced solely from the aligned visual features through an MLP head.

**Image captioning.** We adopt a well-designed fill-in-the-blank prompt template fed into BLIP-Large (Li et al., 2022) to generate captions with images. We design four sentences (Figure 5) to obtain complementary descriptions focusing on the *Simple Description*, *Object Categories*, *Background*, and *Global Visual Complexity*, respectively. For each image in the training set, the prompts are applied sequentially, and the outputs of BLIP are combined to form a complete caption. Each image and its caption constitute an image-text pair, and captions are not re-generated during training. We provide a few examples in Appendix A.8.

**Vision encoder & Text encoder.** The vision encoder in D2S is based on ResNet (He et al., 2016), but the original global pooling and fully connected layers are removed and replaced with a learnable pooling layer, which outputs a one-dimensional feature vector for each image. We employ the pre-trained CLIP (Radford et al., 2021) text encoder as the text encoder in D2S, and use the CLIP tokenizer to segment each caption.

This image shows {...}.  
**Simple Description**  
 The main objects are {...}.  
**Objects Categories**  
 The background looks {...}.  
**Background**  
 The overall visual complexity is {...}.  
**Global Visual Complexity**

Figure 5: Prompt template.

The text encoder remains **frozen during training**, while gradients are updated only for the other components.

**Vision-Text Connector.** Many prior works employ a MLP-style projection layer (Chen et al., 2020) to map visual and textual features into the same dimensional space. Following this practice, we apply **one linear layer** as a projection for vision encoder, aligning it with the dimensionality of the text features. This design ensures that text features can effectively guide visual features within a shared representation space. **Note, the projection will be discarded during inference.**

### 3.2 ENTROPY DISTRIBUTION ALIGNMENT

In the theoretical analysis, we pointed out that the weighted fusion of visual feature entropy  $H_v^F(I)$  and semantic feature entropy  $H_s^F(S)$  can better capture image complexity. However, during model learning, we observe that the empirical distributions of the two modalities show a noticeable bias (Figure 4(d), dotted line). Such bias may introduce additional uncertainty in the fusion stage and weaken the consistency of cross-modal complexity measurement. To address this issue, we propose the entropy distribution alignment. By aligning the entropy distributions of the visual and text modalities, we enhance cross-modal consistency and further improve generalization capability (**Theorem 2**).

**Entropy Buffer.** We establish and maintain two FIFO entropy buffers,  $B_v$  and  $B_s$ , each with capacity  $M$  to store visual feature entropy and textual feature entropy. Before training starts, we duplicate the D2S and freeze it, referring to this frozen copy as the Momentum Model (MoM). At each iteration, we first use the MoM with inference mode to obtain the image and text features of all samples in the mini-batch, and then compute their entropy using Eq.(12). The visual and textual entropy are subsequently stored in their corresponding buffers. Once a buffer reaches half of its capacity, we begin to compute the entropy distribution alignment loss. We update the buffers by adding the new mini-batch samples and removing the oldest entries with same number, thereby reducing distributional bias between old and new members. Furthermore, we update the MoM at each iteration using an exponential moving average (EMA) of gradients (similar as MoCo (He et al., 2020)), and employ the updated MoM to refresh buffer entries with a *refresh step*. This strategy further mitigates distributional drift. We have presented the *Algorithm* of EAL in Appendix A.2.

**Entropy Distribution Alignment Loss  $\mathcal{L}_{\text{eal}}$ .** We define the Entropy Distribution Alignment Loss to encourage consistency between the entropy distributions of visual and textual modalities. Specifically, we adopt the energy distance (Székely & Rizzo, 2013) as our alignment loss. Concretely, let

$$V = \{v_i\}_{i=1}^M, \quad S = \{s_j\}_{j=1}^M \quad (5)$$

denote the entropy values extracted from the image and text modality in the buffers  $B_v$  and  $B_s$ , respectively. The energy distance between the two empirical distributions is defined as

$$D_E^2(V, S) = 2\mathbb{E}[\|V - S\|] - \mathbb{E}[\|V - V'\|] - \mathbb{E}[\|S - S'\|] \quad (6)$$

where,  $V, V'$  are i.i.d. samples from  $B_v$  and  $S, S'$  are i.i.d. samples from  $B_s$ . In practice, the expectations are estimated by averaging over all pairs in a mini-batch:

$$\mathcal{L}_{\text{eal}} = 2 \frac{1}{M^2} \sum_{i=1}^M \sum_{j=1}^M |v_i - s_j| - \frac{1}{M(M-1)} \left( \sum_{i \neq i'} |v_i - v_{i'}| - \sum_{j \neq j'} |s_j - s_{j'}| \right) \quad (7)$$

### 3.3 FEATURE ALIGNMENT

To further exploit textual information, we need to ensure that visual and textual features are aligned in a shared representation space. Inspired by CLIP (Radford et al., 2021), we introduce a contrastive loss that enforces correspondence between visual and textual representations in the joint space. This allows semantic diversity to be effectively injected into visual representations, thereby ensuring  $H_v^F(I) > H_s^F(S)$ . At the same time, the contrastive loss serves as a semantic regularizer, constraining visual features in the shared subspace distribution. This reduces redundancy, lowers the effective dimensionality  $d'$ , and tightens the Rademacher complexity bound, thus improving generalization. Let the outputs of the visual and text encoders be defined as:

$$z_v = f_v(I), \quad z_s = f_s(S) \quad (8)$$

Table 1: **Comparison with state-of-the-art image complexity assessment methods on IC9600.** Entropy buffer size is 2048 with a refresh step of 50 ( $\sim 40$  entropy were refreshed per iteration) and momentum 0.995.  $\lambda$  and  $\gamma$  are set to 5 and 0.01, respectively.  $^\dagger$  denotes an unsupervised method. The results of ICNet and ICCORN are our re-implementation. The **bolded** portion marks the optimal outcome. The underlined portion denotes the second-best.

Method	SRCC	PCC	RMSE	RMAE	Params	Latency
MoCo $^\dagger$ (He et al., 2020)	0.759	0.748	-	-	-	-
SAE $^\dagger$ (Saraee et al., 2020b)	0.865	0.860	0.074	0.240	-	-
CLIC $^\dagger$ (Liu et al., 2024)	0.866	0.858	-	-	-	-
CLICv2 $^\dagger$ (Liu et al., 2025a)	0.879	0.870	-	-	-	-
HyperIQA (Su et al., 2020)	0.935	0.935	0.067	0.229	27.38M	29.638ms
P2P-FM (Ying et al., 2020)	0.940	0.936	0.056	0.208	-	-
TOPIQ (Chen et al., 2024)	0.938	0.944	0.049	-	45.20M	14.434ms
ICNet (Feng et al., 2022)	0.9446	0.9470	0.0582	0.2156	<u>20.33M</u>	<u>9.369ms</u>
std	$\pm 0.0029$	$\pm 0.0011$	$\pm 0.0021$	$\pm 0.0051$		
ICCORN (Guo et al., 2023)	0.9455	0.9490	0.0526	0.2085	22.31M	13.973ms
std	$\pm 0.0011$	$\pm 0.0018$	$\pm 0.0013$	$\pm 0.0053$		
MICM (Li et al., 2025)	0.943	0.953	0.060	-	$\sim 11B$	$\sim 180s$
D2S-R18 (ours)	<u>0.9509</u>	<u>0.9544</u>	<b>0.0495</b>	<b>0.1962</b>	<b>13.02M</b>	<b>4.573ms</b>
std	$\pm 0.0026$	$\pm 0.0009$	$\pm 0.0016$	$\pm 0.0050$		
D2S-R50 (ours)	<b>0.9541</b>	<b>0.9576</b>	<u>0.0496</u>	<u>0.1963</u>	38.72M	10.738ms
std	$\pm 0.0004$	$\pm 0.0010$	$\pm 0.0028$	$\pm 0.0062$		

which are then projected into the joint space through projection layers:

$$\tilde{z}_v = W_v z_v, \quad \tilde{z}_s = W_s z_s \quad (9)$$

where,  $W_v$  and  $W_s$  denote the parameters of the visual and textual projection layers. In this work, we set  $W_s = I$ , meaning that textual features are mapped with an identity transformation, without additional projection. With temperature parameter  $\tau$ , we adopt a InfoNCE loss (van den Oord et al., 2019):

$$\mathcal{L}_{\text{fal}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\tilde{z}_v^i, \tilde{z}_s^i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\tilde{z}_v^i, \tilde{z}_s^j)/\tau)} \quad (10)$$

where,  $\text{sim}(\cdot)$  denotes feature similarity. Positive pairs  $(\tilde{z}_v^i, \tilde{z}_s^i)$  correspond to each image and its caption, while negative pairs  $(\tilde{z}_v^i, \tilde{z}_s^j)$  are drawn from other samples within the mini-batch. The score regression loss adopts mean square error  $\mathcal{L}_{\text{mse}}$ . Finally, the overall training objective of D2S is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda \mathcal{L}_{\text{eal}} + \gamma \mathcal{L}_{\text{fal}} \quad (11)$$

where,  $\lambda$  and  $\gamma$  control the weights of  $\mathcal{L}_{\text{eal}}$  and  $\mathcal{L}_{\text{fal}}$ .

## 4 EXPERIMENTS

The reported results are averaged over three random seeds (42, 826, 1215), and we additionally report the standard deviation (std) in Table 1. More training details are provided in Appendix A.3.

### 4.1 MAIN RESULTS

**Comparison with State-of-the-Art Methods.** We benchmark the proposed D2S model against both unsupervised and supervised approaches on the IC9600 dataset, with results summarized in Table 1. Unsupervised methods exhibit limited performance (best SRCC 0.879). While supervised ones benefit from annotated information, with MICM achieving 0.953 PCC as the previous state-of-the-art, but it requires extremely high computing resources ( $\sim 11B$  params and  $\sim 180s$  latency). Our D2S consistently surpasses these baselines in accuracy, efficiency, and stability. D2S-R18 attains an SRCC of 0.9509 with only 13.02M parameters, reducing inference latency to 4.573ms and

Table 2: **Performance of small samples training on IC9600.** SST@X denotes few-shot training with X samples. Epoch: 5. Hyper-parameters are same as Table 1.

Method	SRCC	PCC	RMSE	RMAE	SRCC	PCC	RMSE	RMAE
	SST@10				SST@50			
ICNet	0.5178	0.5057	0.1527	0.3429	0.5644	0.5864	0.1514	0.3422
ICCORN	0.5221	0.5102	0.1497	0.3393	0.5704	0.5914	0.1484	0.3386
D2S-R18	0.6664	0.6839	0.1302	0.3199	0.8174	0.8332	<b>0.1030</b>	<b>0.2858</b>
D2S-R50	<b>0.6864</b>	<b>0.6972</b>	<b>0.1190</b>	<b>0.3055</b>	<b>0.8515</b>	<b>0.8528</b>	0.1133	0.3025
	SST@100				SST@500			
ICNet	0.8239	0.8259	0.1668	0.3820	0.8942	0.9054	0.0830	0.2582
ICCORN	0.8276	0.8294	0.1605	0.3747	0.8977	0.9080	0.0791	0.2527
D2S-R18	0.8552	0.8547	<b>0.0964</b>	<b>0.2787</b>	0.9115	0.9182	<b>0.0707</b>	<b>0.2326</b>
D2S-R50	<b>0.8680</b>	<b>0.8731</b>	0.1092	0.2913	<b>0.9182</b>	<b>0.9246</b>	0.0709	0.2349

Table 3: **Cross-dataset generalization of D2S on ICA datasets.** Epoch: 5. Hyper-parameters are same as Table 1.

Method	SRCC	PCC	RMSE	RMAE	SRCC	PCC	RMSE	RMAE
	Nagle4k (Nagle & Lavie, 2020b)				VISC-C (Kyle-Davidson et al., 2023)			
ICNet	0.7851	0.7666	0.1082	0.2917	0.7219	0.7111	0.1415	0.3367
ICCORN	0.7905	0.7706	<b>0.1070</b>	<b>0.2891</b>	0.7251	0.7155	0.1407	0.3340
D2S-R18	0.7976	0.7748	0.1102	0.2949	0.7291	0.7165	<b>0.1399</b>	<b>0.3350</b>
D2S-R50	<b>0.7976</b>	<b>0.7765</b>	0.1126	0.2989	<b>0.7317</b>	<b>0.7170</b>	0.1421	0.3375
	Savoias (Saraee et al., 2020b)				VISC-CI (Kyle-Davidson et al., 2023)			
ICNet	0.6813	0.6793	0.1721	0.3736	0.6802	0.6876	<b>0.1549</b>	<b>0.3571</b>
ICCORN	0.6835	0.6811	0.1719	0.3720	0.6825	0.6924	0.1563	0.3591
D2S-R18	0.6780	0.6825	0.1706	0.3717	<b>0.6853</b>	<b>0.6982</b>	0.1580	0.3617
D2S-R50	<b>0.6845</b>	<b>0.6882</b>	<b>0.1700</b>	<b>0.3713</b>	0.6828	0.6963	0.1622	0.3660

outperforming ICCORN at one-fourth the cost. Scaling to ResNet50 further boosts SRCC to 0.9541 and PLCC to 0.9576, setting new state-of-the-art results. Additionally, D2S exhibits markedly lower standard deviations across runs, underscoring its robustness and validating the effectiveness of text-guided visual feature alignment.

**Small Samples Training (SST).** We evaluate D2S under limited data settings by randomly sampling subsets of IC9600 with varying sizes (10, 50, 100, 500), as summarized in Table 2. The results show steady improvements with increasing samples, particularly in SRCC and PCC. Remarkably, with only 10 samples, D2S-R50 achieves an SRCC of 0.6864, far surpassing the ICNet (0.5178), indicating that multi-modal alignment provides semantic constraints that enhance generalization even under extreme data scarcity. Performance rises rapidly with 50–100 samples (e.g., 0.8680 at SST@100) and approaches full-scale results with 500 samples, where both D2S-R18 and D2S-R50 surpass 0.91 in SRCC and PCC while reducing RMSE and RMAE. In contrast, ICNet and ICCORN perform poorly in low-sample regimes. These results demonstrate that D2S not only excels under full supervision but also exhibits strong small-sample learning ability, making it well-suited for real-world scenarios with limited labeled data.

**Cross-Dataset Generalization.** To evaluate generalization, we train D2S on IC9600 and directly test it on Savoias, Nagle4k, and VISC-C/I without fine-tuning (Table 3). Across most metrics on all datasets, both D2S-R18 and D2S-R50 consistently surpass the ICNet and ICCORN, confirming that text-guided alignment enhances robustness under distribution shifts. On Nagle4k and VISC-C, D2S-R50 achieves SRCC values of 0.7976 and 0.7317, outperforming other methods and demonstrating the transferability of cross-modal modeling. Although correlations on Savoias and VISC-C/I remain below 0.70, D2S still matches or slightly exceeds others. The marginal gap between ResNet18 and

Table 4: **Cross-task transfer of D2S to NR-IQA.** D2S were trained for 5 epochs with input resolution 384. The method references and implementation details are provided in the Appendix A.3.

Method	Source	KADID-10K		KonIQ-10K		TID2013	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
QPT	CVPR 2023	0.925	0.928	-	-	0.895	0.914
ARNIQA	WACV 2024	0.908	0.912	-	-	0.880	0.901
TOPIQ	TIP 2024	0.921	0.924	0.574	0.657	0.870	0.884
CDINet	TMM 2024	0.920	0.919	0.865	0.880	0.898	0.908
LoDa	CVPR 2024	0.876	0.899	0.932	0.944	0.869	0.901
ADTRS	ICIP 2024	-	-	0.905	0.918	0.878	0.897
VISGA	TCSVT 2025	0.919	0.925	0.930	0.937	0.901	0.914
CoDI-IQA	ArXiv 2025	0.936	0.940	0.902	0.917	0.901	0.916
DGIQA	ArXiv 2025	0.943	0.945	<b>0.934</b>	<b>0.942</b>	0.934	0.940
RSFIQA	ArXiv 2025	<u>0.953</u>	<u>0.954</u>	<u>0.934</u>	<u>0.940</u>	<b>0.951</b>	<b>0.959</b>
D2S-R18	ours	0.952	0.953	0.901	0.922	<u>0.941</u>	<u>0.938</u>
D2S-R50		<b>0.958</b>	<b>0.959</b>	0.900	0.925	0.938	0.935

Table 5: **Ablation study of the main components in D2S on IC9600.** Each case (a ~ e) corresponds to different combinations of these modules.

Case	AttnPool	EAL	FAL	SRCC	PCC	RMSE	RMAE
(a)	×	×	×	0.9396	0.9430	0.0547	0.2052
(b)	✓	×	×	0.9446	0.9476	0.0541	0.2050
(c)	✓	✓	×	0.9467	0.9503	0.0546	0.207
(d)	✓	×	✓	0.9473	0.9510	0.0551	0.2086
(e)	✓	✓	✓	<b>0.9499</b>	<b>0.9540</b>	<b>0.0472</b>	<b>0.1906</b>

ResNet50 suggests that gains stem mainly from the alignment mechanism rather than backbone scale. These results highlight the effectiveness of semantic-driven complexity modeling in achieving consistent cross-dataset generalization.

**Cross-Task Transfer.** We further test the cross-task generalization of D2S by transferring it to NR-IQA, where all baseline results are taken from original papers. Training and Datasets details are provide in Appendix A.3 and A.4. D2S achieves performance comparable to or surpassing recent methods. On KADID-10K, D2S-R50 attains 0.958 SRCC and 0.959 PLCC, outperforming all existing approaches. On KonIQ-10K, although DGIQA (0.934 SRCC) remains superior, D2S still yields competitive results around 0.90 SRCC. On TID2013, both variants exceed 0.938, again most existing approaches. These findings confirm that semantic alignment not only enhances image complexity assessment but also transfers effectively to perceptual quality assessment.

## 4.2 ANALYSIS OF D2S

**Ablation study of the main components.** To evaluate the contributions of each module in D2S, we gradually introduced AttnPool, EAL, and FAL, and the results are summarized in Table 5. The introduction of AttnPool significantly improved the performance of SRCC and PCC, as it can achieve fine visual aggregation. FAL can generate slightly higher correlation, while EAL achieves better stability by reducing errors. When these three components are combined, D2S achieved the best overall results (SRCC 0.9499, PCC 0.9540), and there was a significant reduction in RMSE and RMAE. These findings confirm that these modules are complementary, and their integration provides balanced improvements in accuracy and robustness.

**What Does D2S Learn?** We analyzed the top-20 most activated channels (Figure 7) across IC9600 and found that D2S relies on a sparse subset of discriminative channels rather than distributing attention evenly. Importantly, these activations show no strong correlation with low-level statistics



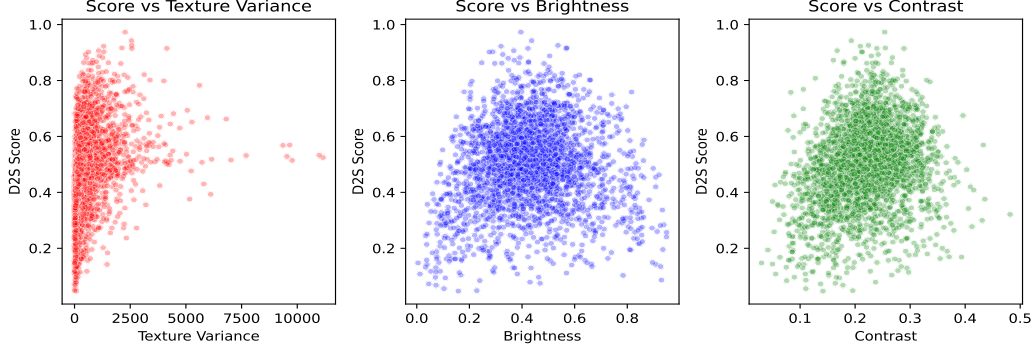


Figure 6: The correlation between the predicted scores of D2S and the low-level statistics, including texture variance, brightness, and contrast.

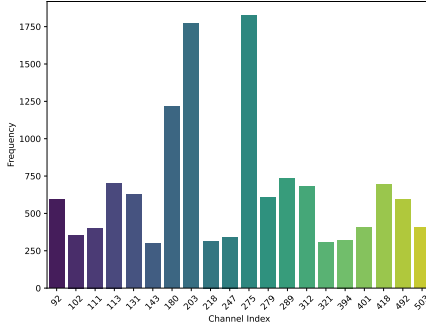


Figure 7: **Channel utilization histogram.** Frequencies of the top-20 most correlated feature channels across the IC9600 test set.

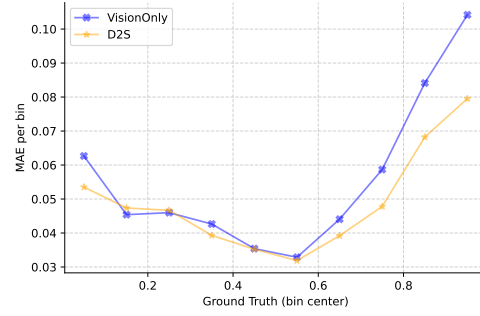


Figure 8: **Binned MAE comparison.** Mean absolute error (MAE) across 10 uniformly spaced complexity bins.

such as texture, brightness, or contrast (Figure 6), suggesting that the model captures transferable semantic or structural patterns beyond superficial cues.

**Binned Error Analysis.** We further divided predicted scores into 10 bins to error distribution (Figure 8). D2S performs on par with the visual-only baseline in simple scenes (complexity  $< 0.55$ ), but consistently outperforms it as complexity increases, with the gap widening in high-complexity cases. This demonstrates that text-guided alignment is particularly valuable for modeling intricate scenes where visual cues alone are insufficient. **Further analyses and ablations are in Appendix A.6, with discussions in A.9 and limitations in A.10.**

## 5 CONCLUSION

In this work, we addressed the challenge of image complexity assessment by introducing multi-modal fusion into complexity modeling. We proposed the D2S framework, which leverages pre-trained VLMs to describe images and integrates feature alignment and entropy alignment mechanisms to guide complexity assessment. Our theoretical analysis demonstrated that combining visual and semantic features enriches representation diversity and reduces the effective hypothesis space, thereby improving both accuracy and generalization. Extensive experiments validated these insights: D2S not only outperformed state-of-the-art methods on the IC9600 benchmark but also showed competitive transferability on image quality assessment datasets. Furthermore, the framework achieves this without incurring additional multi-modal inference cost, as only the visual branch is required at test time. Taken together, these results highlight the effectiveness and efficiency of semantic align-

ment for complexity modeling and point toward the broader potential of multi-modal integration in perceptual understanding tasks.

## REFERENCES

- Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.
- Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arnika: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024.
- Mohammed Alsaafin, Musab Alsheikh, Saeed Anwar, and Muhammad Usman. Attention down-sampling transformer, relative ranking and self-consistency for blind image quality assessment. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 1260–1266. IEEE, 2024.
- Luigi Celona, Gianluigi Ciocca, and Raimondo Schettini. On the use of visual transformer for image complexity assessment. *Proceedings Copyright*, 640:647, 2024.
- Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmlR, 2020.
- Yan-Qin Chen, Jin Duan, Yong Zhu, Xiao-Fei Qian, and Bo Xiao. Research on the image complexity based on neural network. In *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pp. 295–300. IEEE, 2015.
- Valeriy Chikhman, Valeriya Bondarko, Marina Danilova, Anna Goluzina, and Yuri Shelepin. Complexity of images: Experimental and computational estimates compared. *Perception*, 41(6):631–647, 2012.
- Susan F Chipman. Complexity and structure in visual patterns. *Journal of Experimental Psychology: General*, 106(3):269, 1977.
- Lingchen Dai, Kang Zhang, Xianjun Sam Zheng, Ralph R Martin, Yina Li, and Jinhui Yu. Visual complexity of shapes: a hierarchical perceptual learning model. *The Visual Computer*, 38(2): 419–432, 2022.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- T Feng, Y Zhai, J Yang, J Liang, DP Fan, J Zhang, L Shao, and D Tao. Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8577–8593, 2022.
- Alexandra Forsythe. Visual complexity: Is that all there is? In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pp. 158–166. Springer, 2009.
- Xiaoying Guo, Lu Wang, Tao Yan, and Yanfeng Wei. Image visual complexity evaluation based on deep ordinal regression. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 199–210. Springer, 2023.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/jia21b.html>.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.
- Cameron Kyle-Davidson, Adrian G Bors, and Karla K Evans. Predicting human perception of scene complexity. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 1281–1285. IEEE, 2022.
- Cameron Kyle-Davidson, Elizabeth Yue Zhou, Dirk B Walther, Adrian G Bors, and Karla K Evans. Characterising and dissecting human perception of scene complexity. *Cognition*, 231:105319, 2023.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.
- Pu Li, Yi Yang, Wangda Zhao, and Miao Zhang. Evaluation of image fire detection algorithms based on image complexity. *Fire safety journal*, 121:103306, 2021.
- Yixiao Li, Xiaoyuan Yang, Yuqing Luo, Hadi Amirpour, Hantao Liu, and Wei Zhou. Unlocking implicit motion for evaluating image complexity. *Displays*, 90:103131, 2025.
- Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.
- Shipeng Liu, Liang Zhao, and Dengfeng Chen. Clic: Contrastive learning framework for unsupervised image complexity representation, 2024. URL <https://arxiv.org/abs/2411.12792>.
- Shipeng Liu, Liang Zhao, and Dengfeng Chen. Clicv2: Image complexity representation via content invariance contrastive learning, 2025a. URL <https://arxiv.org/abs/2503.06641>.
- Shuai Liu, Qingyu Mao, Chao Li, Jiacong Chen, Fanyang Meng, Yonghong Tian, and Yongsheng Liang. Content-distortion high-order interaction for blind image quality assessment, 2025b. URL <https://arxiv.org/abs/2504.05076>.
- Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- Penousal Machado, Juan Romero, Marcos Nadal, Antonino Santos, João Correia, and Adrián Carballal. Computerized measures of visual complexity. *Acta psychologica*, 160:43–57, 2015.
- Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.
- Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.
- Fintan Nagle and Nilli Lavie. Predicting human complexity perception of real-world scenes. *Royal Society open science*, 7(5):191487, 2020a.

- Fintan Nagle and Nilli Lavie. Predicting human complexity perception of real-world scenes. *Royal Society open science*, 7(5):191487, 2020b.
- Aude Olivia, Michael L Mack, Mochan Shrestha, and Angela Peeper. Identifying the perceptual dimensions of visual complexity of scenes. In *Proceedings of the annual meeting of the cognitive science society*, volume 26, 2004.
- Letizia Palumbo, Ruth Ogden, Alexis DJ Makin, and Marco Bertamini. Examining visual complexity and its influence on perceived duration. *Journal of vision*, 14(14):3–3, 2014.
- Sang-Min Park and Young-Gab Kim. Visual language integration: A survey and open challenges. *Computer Science Review*, 48:100548, 2023.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- Vaishnav Ramesh, Junliang Liu, Haining Wang, and Md Jahidul Islam. Dgiqua: Depth-guided feature attention and refinement for generalizable image quality assessment. *arXiv preprint arXiv:2505.24002*, 2025.
- Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195:102949, 2020a.
- Elham Saraee, Mona Jalal, and Margrit Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195:102949, 2020b.
- Tingke Shen, Surabhi S Nath, Aenne Brielmann, and Peter Dayan. Simplicity in complexity: Explaining visual complexity using deep segmentation models. *arXiv preprint arXiv:2403.03134*, 2024.
- Haozhi Shi, Weiyang Xie, Haonan Qin, Yunsong Li, and Leyuan Fang. Visual state space model with graph-based feature aggregation for blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- Sandhya Singh and Dolley Shukla. A review on various measures for finding image complexity. *International Journal of Scientific Research Engineering & Technology*. ISSN, pp. 2278–0882, 2017.
- Chenyue Song, Chen Hui, Haiqi Zhu, Feng Jiang, Yachun Mi, Wei Zhang, and Shaohui Liu. Segmenting and understanding: Region-aware semantic attention for fine-grained image quality assessment with large language models, 2025. URL <https://arxiv.org/abs/2508.07818>.
- Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3667–3676, 2020.
- Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL <https://arxiv.org/abs/1807.03748>.
- Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 2025.
- Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019.

- Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.
- Kangmin Xu, Liang Liao, Jing Xiao, Chaofeng Chen, Haoning Wu, Qiong Yan, and Weisi Lin. Boosting image quality assessment through efficient transformer adaptation with local feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2662–2672, 2024.
- Miin-Shen Yang and Yessica Nataliani. A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy. *IEEE Transactions on Fuzzy Systems*, 26(2):817–835, 2017.
- Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3575–3585, 2020.
- Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.
- Kai Zhao, Kun Yuan, Ming Sun, Mading Li, and Xing Wen. Quality-aware pre-trained models for blind image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22302–22313, 2023.
- Limin Zheng, Yu Luo, Zihan Zhou, Jie Ling, and Guanghui Yue. Cdinet: Content distortion interaction network for blind image quality assessment. *IEEE Transactions on Multimedia*, 26: 7089–7100, 2024.

## A APPENDIX

### A.1 PROOF OF PROPOSITION 1

**Proof.** We first define the entropy of visual features and semantic features. Assume the visual encoder outputs a feature distribution  $\{p_k^{(v)}\}_{k=1}^K$ , and the text encoder outputs a feature distribution  $\{p_t^{(s)}\}_{t=1}^T$ . Their corresponding entropy is defined as:

$$H_v^F(I) = -\sum_{k=1}^K p_k^{(v)} \log p_k^{(v)}, \quad H_s^F(S) = -\sum_{t=1}^T p_t^{(s)} \log p_t^{(s)} \quad (12)$$

where,  $H_v^F(I)$  measures the diversity of the visual space (e.g., texture, color, structure), while  $H_s^F(S)$  measures the diversity of the semantic space (e.g., object categories, relationships, actions). We then consider multi-modal fusion. Since multi-modal feature space can simultaneously contain both visual and semantic diversity, we define the fused entropy as:

$$H^F(I) = \alpha H_v^F(I) + \beta H_s^F(S) \quad (13)$$

Obviously, when  $\beta > 0$ , we have  $H^F(I) > H_v^F(I)$ . On the other hand, since feature entropy is positively correlated with the original entropy, i.e.,  $H^F(I) \propto H(I)$ , and entropy reflects image complexity to some extent, i.e.  $H \propto C$ , we can conclude that image-text fusion leads to results closer to the true image complexity:

$$C_v(I) < \alpha C_v(I) + \beta C_s(S) \propto C(I) \quad (14)$$

where,  $C_v(I)$  denotes the complexity contribution from the visual branch, and  $C_s(S)$  denotes the complexity contribution from the text branch. This implies that the complexity obtained after multi-modal fusion is closer to the true complexity of the image.

### A.2 EAL ALGORITHM

---

**Algorithm 1:** Entropy Distribution Alignment

---

**Input:** Image  $I$ , Caption  $S$ , D2S  $f_\theta$ , MoM  $f_\xi$ , Momentum  $m$ , FIFO Buffers  $B_v, B_s$ , Buffer size  $M$ , Refresh step  $r$

**Before training:** Create  $f_\theta$ , copy as  $f_\xi$ ; Create  $B_v, B_s$  with  $M$ .

**for each mini-batch do**

```

     $f_\xi$  extract features  $z_v, z_s$  via Eq.(8);
    Compute entropy  $H_v, H_s$  via Eq.(12);
    Store into buffers  $B_v$  and  $B_s$ , meanwhile remove oldest ones;
    if  $B_v$  and  $B_s$  Entries number  $\geq \frac{M}{2}$  then
        | Compute  $\mathcal{L}_{\text{eal}}$  from  $B_v$  and  $B_s$ ;
    else
        |  $\mathcal{L}_{\text{eal}} = 0$ 
    Update  $f_\theta$  via gradient;
    Update  $f_\xi$  via  $\xi \leftarrow m\xi + (1 - m)\theta$ ;
    Refresh  $r$  old entries via updated  $f_\xi$ ;

```

---

### A.3 IMPLEMENTATION DETAILS

We implement our model in PyTorch. The ResNet backbone is initialized with ImageNet-pretrained (Deng et al., 2009) weights from TIMM (Wightman, 2019), and the BLIP caption generator is employed in its large configuration. The CLIP text encoder is also initialized with pretrained parameters. We train the model using the Adam (Adam et al., 2014) optimizer with weight decay  $1e-3$ . The initial learning rate was set to  $1e-3$  and the batch size to 32; cosine annealing (Loshchilov & Hutter, 2016) was used for scheduling, with a minimum learning rate of  $2.5e-6$ . The temperature was set to 0.07. Training was performed on a single NVIDIA RTX 3090.

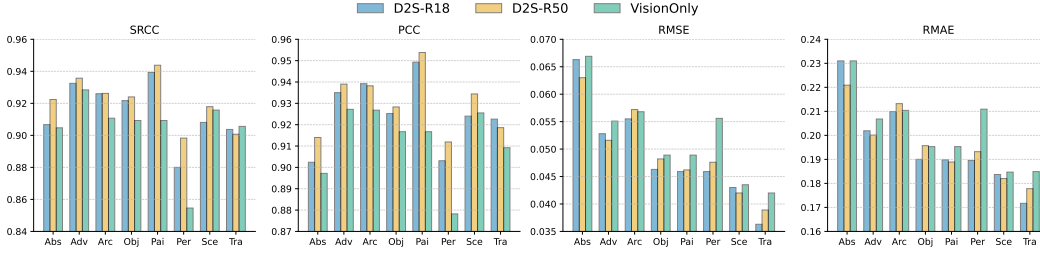


Figure 9: Performance of semantic categories in IC9600.

For the results reported in Table 2, the training set was constructed by dividing the complexity score range (0, 1) into ten intervals and randomly sampling within each interval. Specifically, the SST@10 setting samples one image from each interval, whereas SST@50 samples five images from each interval. It is worth noting that, due to the scarcity of high-complexity images, the number of intervals containing samples in SST@500 is fewer than fifty. Nevertheless, we refer to this configuration as SST@500 for consistency.

For the results in Table 3, we adopted a cross-dataset evaluation protocol, where the full datasets from the ICA benchmark, including Nagle4k (Nagle & Lavie, 2020b), Savoias (Saraee et al., 2020b), and VISC-C/I (Kyle-Davidson et al., 2023), were used as test sets.

For the results in Table 4, we followed the common evaluation protocol in NR-IQA. After min-max normalizing the MOS scores of all images, each dataset was randomly partitioned into training, validation, and test subsets with a ratio of 6:2:2. We compared D2S with QPT (Zhao et al., 2023), ARNIQA (Agnolucci et al., 2024), TOPIQ (Chen et al., 2024), CDINet (Zheng et al., 2024), LoDa (Xu et al., 2024), ADTRS (Alsaafin et al., 2024), VISGA (Shi et al., 2025), CoDI-IQA (Liu et al., 2025b), DGIQA (Ramesh et al., 2025) and RSFIQA (Song et al., 2025).

#### A.4 DATASETS

The datasets used in our study are drawn from two tasks: image complexity assessment (ICA) and image quality assessment (IQA). The ICA datasets include IC9600, Savoias, PASCAL VOC\_4000 (Nagle4k), and VISC-C/I. Among them, IC9600 is employed to validate the effectiveness of our method, while Savoias, Nagle4k, and VISC-C/I are used to evaluate cross-dataset generalization. The IQA datasets, consisting of KADID-10K, KonIQ-10K, and TID2013, are utilized to assess the cross-task transferability of the proposed approach.

#### A.5 EVALUATION METRICS

We adopt Spearman’s Rank Correlation Coefficient (SRCC), Pearson’s Linear Correlation Coefficient (PLCC), Root Mean Square Error (RMSE), and Relative Mean Absolute Error (RMAE) to evaluate the predictive performance of our model. In addition, we report the number of parameters (Params) and inference latency (Latency) to assess computational efficiency.

#### A.6 FURTHER ANALYSES AND ABLATIONS

**Performance of semantic category in IC9600.** To analyze the impact of semantic alignment, we evaluate D2S-R18, D2S-R50, and the VisionOnly variant across semantic categories in IC9600, as shown in Figure 9. Both D2S models consistently surpass VisionOnly in correlation (SRCC/PCC) and error metrics (RMSE/RMAE), with notable gains in object-centric (*Obj*, *Per*) as well as abstract and artistic categories (*Abs*, *Pai*), where visual cues alone are insufficient. D2S-R50 further outperforms D2S-R18, particularly in *Arc* and *Sce*, indicating that larger backbones better exploit semantic features. Significant error reductions in categories such as *Tra* and *Sce* highlight the benefits of semantic guidance in complex or cluttered scenes. Overall, these results confirm that feature alignment enhances not only average accuracy but also robustness across diverse semantic distributions.

Table 6: **Ablation study of EAL hyper-parameters on IC9600.**  $M$ : buffer size, tested with 512, 1024, 2048, 4096 entries. **mom.**: momentum for the MoM update, tested with 0.990, 0.995, 0.999, 0.9999. **steps**: buffer refresh step, tested with 16, 50, 128 iterations, corresponding to the number of entries updated per iteration (128, 40, 16, respectively).

$M$	mom.	steps	SRCC	PCC	RMSE	RMAE
2048	0.99	50	0.9487	0.9533	0.0497	0.1955
	0.995		<b>0.9508</b>	<b>0.9545</b>	<b>0.0496</b>	<b>0.1962</b>
	0.999		0.9499	0.9540	0.0472	0.1906
	0.9999		0.9484	0.9527	0.0497	0.1960
512	0.995	50	0.9494	0.9531	0.0519	0.2016
1024			0.9493	0.9533	0.0516	0.2008
2048			<b>0.9499</b>	<b>0.9540</b>	<b>0.0472</b>	<b>0.1906</b>
4096			0.9501	0.9530	0.0529	0.2041
2048	0.995	16	0.9499	0.9539	0.0513	0.2008
		50	<b>0.9499</b>	<b>0.9540</b>	<b>0.0472</b>	<b>0.1906</b>
		128	0.9496	0.9539	0.0502	0.1973

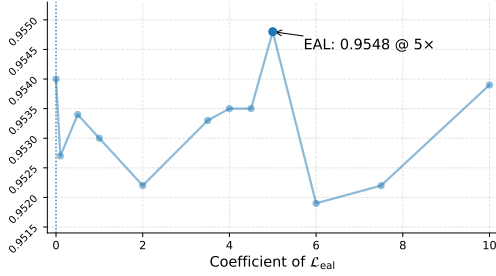


Figure 10: **EAL coefficient ablation.** D2S is trained for 20 epochs with ResNet18 as the backbone, while fixing the FAL coefficient at 0.01. The best result is achieved when the EAL coefficient equals 5.

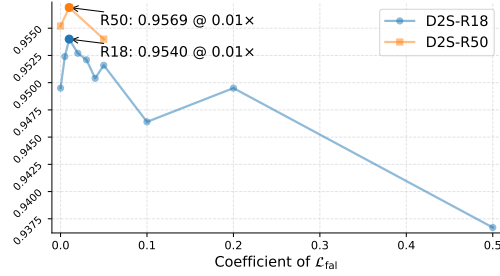


Figure 11: **FAL coefficient ablation.** D2S is trained for 20 epochs with ResNet18 and ResNet50 as backbones, while fixing the EAL coefficient at 0.01. The best result is achieved when the FAL coefficient equals 0.01.

**EAL ablation.** We fix ResNet18 as the backbone in D2S and set the FAL coefficient to 0.01, the buffer size to 2048, the momentum to 0.995, and the refresh steps to 50. By varying the EAL coefficient, we identify the optimal setting, as shown in Figure 10. The best result is obtained when the EAL coefficient equals 5, achieving a PCC of 0.9548. We also conduct additional ablation studies on other hyper-parameters related to EAL, and the results are summarized in Table 6. The best configurations are obtained with a buffer size of 2048, a momentum of 0.995, and a refresh steps of 50.

**FAL coefficient ablation.** We also investigate the effect of the Feature Alignment Loss (FAL) by varying its coefficient. The hyper-parameters of EAL are fixed at best configuration. Experiments are conducted with both ResNet18 and ResNet50 as backbones. As shown in Figure 11, the best results are achieved when the FAL coefficient is set to 0.01.

**Impact of Different Caption Generators.** We investigate how the choice of caption generator affects the performance of D2S. Specifically, we replace the BLIP captions with those generated by Florence-2 (Xiao et al., 2024). We generate captions by Florence-2 for CapI <prompt:caption>, CapII <prompt:detailed caption>, and CapIII <prompt:more detailed caption>. The results are summarized in Table 7 right part. Overall, D2S exhibits stable performance across different caption sources, with minor variations in SRCC and PCC. Notably, using BLIP captions achieves the best performance (SRCC 0.9476, PCC 0.9524), suggesting that high-quality, descriptive captions can better guide the modeling of semantic information for complexity assessment.



Table 7: **Comparisons with different captions.** ‘Only Caption’ indicates using captions alone for IC evaluation. ‘Image & Caption Concat’ means concatenating the image features and text features before inputting them into the regression head.

CapType	Only Caption				Image & Caption Concat			
	SRCC	PCC	RMSE	RMAE	SRCC	PCC	RMSE	RMAE
CapI	0.7095	0.7162	0.1084	0.2868	0.9456	0.9501	0.0535	0.2044
CapII	0.8120	0.8172	0.0903	0.2630	0.9469	0.9506	0.0534	0.2044
CapIII	0.8102	0.8151	0.0904	0.2626	0.9441	0.9484	0.0524	<b>0.2015</b>
BLIP	<b>0.8260</b>	<b>0.8251</b>	<b>0.0888</b>	<b>0.2609</b>	<b>0.9476</b>	<b>0.9524</b>	<b>0.0524</b>	0.2029

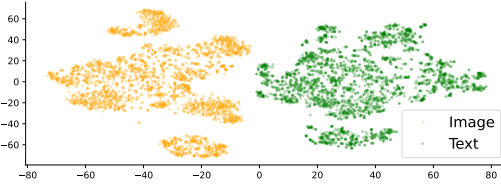


Figure 12:  $t$ -SNE before alignment

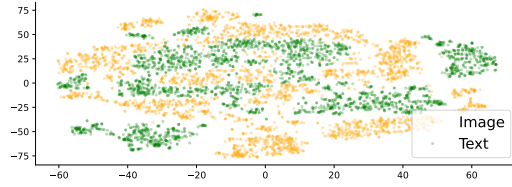


Figure 13:  $t$ -SNE after alignment

**Performance Using Only Captions.** To further understand the contribution of semantic information, we evaluate models trained solely on textual descriptions without visual input. As shown in Table 7 left part, the performance is naturally lower than the full D2S model, with SRCC ranging from 0.7095 (CapI) to 0.8260 (BLIP). This confirms that textual features alone carry informative cues about image complexity but are insufficient to fully capture visual diversity.

**Visualization of Modal Alignment.** To better understand the effect of the alignment mechanism, we visualize the distributions of visual and textual features using  $t$ -SNE. In Figure 12, before alignment, the two modalities are clearly separated, with visual features and textual features clustered in distinct regions. In Figure 13, after alignment, the two modalities become interleaved, indicating that the alignment module effectively bridges the semantic gap and encourages cross-modal consistency.

In addition, we further examine the entropy distributions of visual and textual modalities. As shown in Figure 14, before alignment (dashed lines without fill), the overlap between visual and textual entropy distributions is relatively limited. After alignment (solid lines with fill), the overlap increases substantially, suggesting that the alignment mechanism not only brings the feature embeddings closer in the shared space but also enhances the consistency of their statistical properties. Together, these visualizations provide intuitive evidence that semantic alignment enables the model to construct a unified representation space, which facilitates more accurate complexity assessment.

**Prediction Scatter and Regression Line.** We also analyze the regression behavior by plotting predicted scores against ground-truth labels, along with the ideal regression line. As illustrated in Figure 15, the predictions of D2S are much closer to the ideal line compared to the unimodal baseline. This indicates that D2S not only achieves higher correlation with ground-truth labels but also reduces systematic prediction bias. The tighter distribution around the regression line further confirms the stability and accuracy of the proposed approach.

**Feature Activation Visualization.** We further investigate how D2S attends to image regions by visualizing different types of feature activation maps. Specifically, we compare the channel-wise maximum activation, the channel-wise mean activation, and Grad-CAM responses. As shown in Figure 16, the mean activation map highlights the largest number of relevant regions, suggesting that averaging across channels provides a more comprehensive representation of structural complexity. The maximum activation map illuminates fewer areas but still captures several salient regions. In contrast, Grad-CAM produces relatively sparse responses, indicating that conventional gradient-based attention may not fully capture the fine-grained complexity cues. These observations demonstrate that D2S benefits from richer and more distributed feature aggregation.

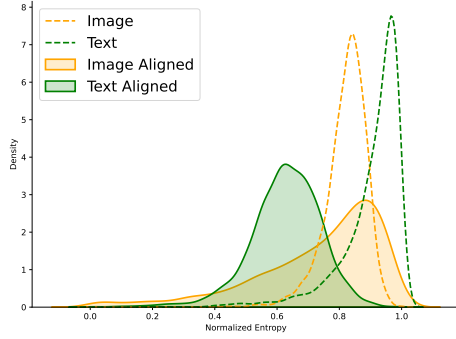


Figure 14: **Entropy distribution alignment.** Entropy distributions of visual and textual modalities before (dashed lines) and after alignment (solid lines).

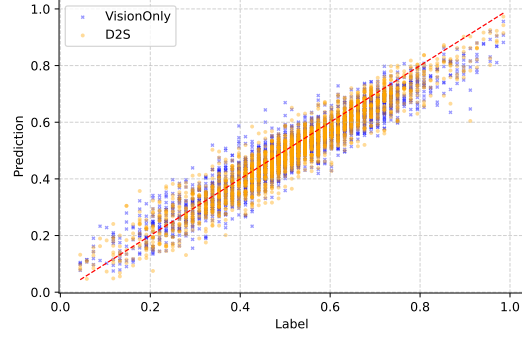


Figure 15: **Prediction scatter plot.** Scatter plots of predicted scores versus ground-truth labels.

## A.7 RELATED WORKS

### A.7.1 IMAGE COMPLEXITY ASSESSMENT.

**Statistical features.** Early methods for image complexity assessment primarily relied on statistical features or low-level visual indicators, such as entropy (Chikhman et al., 2012), symmetry (Chipman, 1977; Kyle-Davidson et al., 2022), spatial layout (Olivia et al., 2004), and compressibility (Palumbo et al., 2014; Machado et al., 2015). These approaches are easy to implement and highly interpretable, but they suffer from clear limitations. Specifically, they mainly capture local structures and texture details, while being sensitive to noise and image resolution.

**Deep learning models.** With the advent of deep learning, several approaches attempted to directly model image complexity using learning-based frameworks (Nagle & Lavie, 2020a; Saraee et al., 2020a; Kyle-Davidson et al., 2022). For instance, ICNet (Feng et al., 2022) improves complexity regression by combining multi-scale inputs with convolutional features, while ICCORN (Guo et al., 2023) employs a larger backbone and integrates ordinal regression constraints to enhance perceptual modeling. Celona et al. (Celona et al., 2024) further introduced ViTs for complexity assessment, highlighting the potential of deep features in complexity modeling. Nevertheless, these models largely focus on low-level image cues related to complexity, overlooking the human tendency to rely on high-level semantic information when making judgments.

**High-level semantics.** To our knowledge, Shen et al. (Shen et al., 2024) were the first to introduce high-level semantics into this field. They quantified the number of segments and categories in an image using SAM (Kirillov et al., 2023) and FC-CLIP (Yu et al., 2023), respectively, and employed a linear regression model to predict complexity scores. This improved interpretability, but yielded suboptimal performance. Li et al. (Li et al., 2025) argued that implicit motion in image objects could benefit ICA, leveraging VLMs to generate simple captions that served as prompts to convert static images into dynamic videos. By fusing video, image, and text branches, they achieved the best performance to date on IC9600, but at the cost of  $\sim 11$ B parameters and substantial computational resources.

**Our approach.** In contrast, our approach extracts high-level semantic information from VLMs through carefully designed prompt templates, using it only to guide the image branch during training. **At inference, our model requires only the visual input.**

### A.7.2 VISION-LANGUAGE MODELING.

In recent years, visual-language models (VLMs) have achieved remarkable progress in tasks such as image captioning (Li et al., 2022), cross-modal retrieval (Wang et al., 2025), and visual question answering (Kuang et al., 2025), with representative models including CLIP, BLIP, and ALIGN (Jia et al., 2021). By aligning vision and language representations, these models effectively capture the

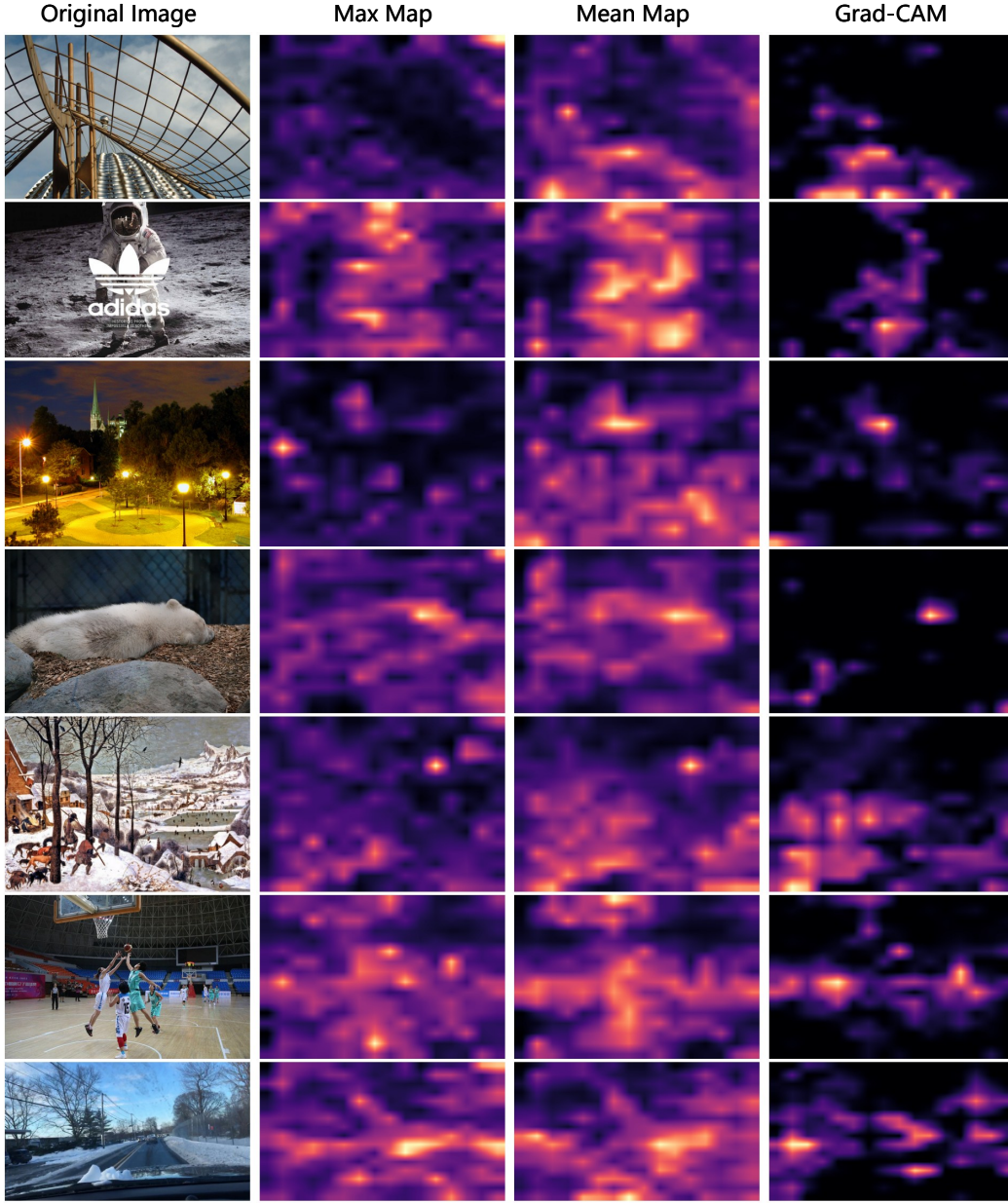


Figure 16: **Feature activation maps.** Visualization of channel activation maps (stage 4 in ResNet) using three approaches: maximum channel activation, mean channel activation, and Grad-CAM.

semantic information of objects, relationships, and scenes. Prior studies show that language descriptions provide complementary information beyond low-level features, enhancing a model’s ability to understand and quantify image content. However, research on visual-language alignment mechanisms for image complexity modeling remains limited. Current complexity assessment methods have not fully exploited semantic information to improve cross-domain generalization. This gap motivates our Describe-to-Score framework, which achieves unified modeling of low-level vision and high-level semantics through visual-language fusion, thereby enhancing both the accuracy and generalization of complexity assessment.

Table 8: Comparison of complexity predictions for different modalities. All the results are derived from the aforementioned table (rounded to three decimal places).

Method	Modal	SRCC	PCC	RMSE	RMAE
CapI	Text-only	0.710	0.716	0.108	0.287
BLIP		0.826	0.825	0.089	0.261
CLICv2	Vision-only	0.879	0.870	-	-
HyperIQA		0.935	0.935	0.067	0.229
ICNet		0.945	0.947	0.058	0.216
ICCORN		0.946	0.949	0.053	0.209
MICM	Vision-Text fusion	0.943	0.953	0.060	-
D2S-R18		0.951	0.954	0.050	0.196
D2S-R50		<b>0.954</b>	<b>0.958</b>	<b>0.050</b>	<b>0.196</b>

#### A.8 IMAGE CAPTION EXAMPLES

In Figure 17 and 18, we present some examples of captions. We selected one image-caption pair from each of the ten score ranges. The prompt template text is marked in green, while the generated text is marked in black or red. We found that the last generated text was more incorrect (in red). We believe this might be due to the overly abstract language. VLMs are unable to generate features in the image that match them, resulting in unexpected answer.

#### A.9 DISCUSSION

##### Does ICA really require high-level semantic information?

A fundamental question in this work is whether high-level semantic information is necessary for accurate image complexity assessment. Low-level visual cues such as texture, color, and edge density provide a baseline measure of complexity, but human perception may also rely on semantic content, which can potentially improve prediction accuracy.

Table 8 compares the performance of multiple methods across three modalities: Text-only (CapI, BLIP), Vision-only (CLICv2, HyperIQA, ICNet, ICCORN), and Vision-Text fusion (MICM, D2S R18/R50). Text-only methods rely purely on image captions. CapI achieves PCC 0.716 and BLIP improves to 0.825, indicating that textual cues capture part of the complexity information but are insufficient for high-precision prediction. Vision-only methods leverage visual features, with unsupervised methods like CLICv2 (PCC 0.870) learning content-invariant complexity representations. Supervised visual models reach PCC 0.949, highlighting that low-level visual features dominate complexity perception and provide strong baseline performance. Vision-Text fusion methods integrate both visual and semantic cues. MICM achieves PCC 0.953, and D2S-R50 reaches 0.958, with corresponding RMSE and RMAE showing the lowest prediction errors. Compared to vision-only baselines, this demonstrates a clear improvement, suggesting that semantic guidance provides measurable benefits in complexity assessment.

**Observations from these results.** Textual information alone is limited. While caption-based models partially capture complexity cues (PCC 0.716–0.825), they cannot match vision-based performance. Vision dominates baseline performance. Supervised vision-only models already achieve high PCC (0.947–0.949), confirming that low-level visual features are sufficient for most complexity signals. Semantic cues refine predictions. Vision-text fusion models consistently outperform vision-only methods, albeit the improvement is smaller than the jump from text-only to vision-only. This indicates that high-level semantics are not strictly necessary but provide a subtle enhancement, particularly for fine-grained correlation with human perception.

Therefore, the progression from Text-only to Vision-only to Vision-Text fusion highlights that while low-level visual features are the primary driver of complexity prediction, high-level semantic information can further refine accuracy. D2S exemplifies this synergy, leveraging semantic cues to enhance performance without replacing the foundational value of visual representation.

#### A.10 LIMITATIONS

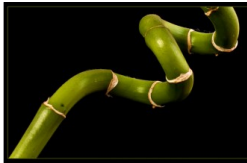
Despite the promising results, our work has several limitations. First, the annotation of image complexity is inherently subjective and the available datasets may not fully cover all domains. Second, our theoretical analysis relies on assumptions about entropy–complexity correspondence and semantic regularization, which may not hold in extreme or synthetic cases. Third, while the CLIP-style alignment loss proves effective, we have not explored alternative contrastive formulations or larger-scale pretrained backbones. Finally, our evaluations mainly focus on ICA/IQA benchmarks, leaving open questions about performance in other modalities and more diverse downstream tasks. We leave these directions for future research.





**Score: 0.05882352963089943**

*this image shows* a lego brick on a white background. *the main objects are* arranged in a single line on the white background. *the background looks* like a lego building with a small window. *the overall visual complexity is* a bit of a lego.



**Score: 0.1617647111415863**

*this image shows* a close up of a green plant with a black background. *the main objects are* green and the stems are brown. *the background looks* like a black background with a picture of a green plant. *the overall visual complexity is* very similar to the bamboo plant.



**Score: 0.25**

*this image shows* a person 's hands in a box with a hole in it. *the main objects are* in a box with two hands. *the background looks* black and dark, but the hands are visible. *the overall visual complexity is* a very important feature in this photograph.



**Score: 0.3529411852359772**

*this image shows* a picture of a baseball game with the name of the team. *the main objects are* in the game and the text is in the background. *the background looks* like a fire with a baseball and a ball. *the overall visual complexity is* clearly visible in the title screen of the game.



**Score: 0.45588234066963196**

*this image shows* a painting of two women sitting on a chair. *the main objects are* in the painting. *the background looks* like a painting of two women sitting in a chair. *the overall visual complexity is* very similar to the painting above.



**Score: 0.5588235259056091**

*this image shows* a tall building with many windows and wires. *the main objects are* reflected in the windows of the building. *the background looks* like a reflection of a building in a mirror. *the overall visual complexity is* clearly visible in this picture.

Figure 17: Image-caption pair examples (part 1).



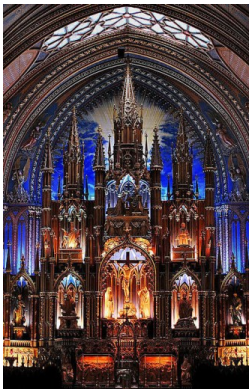
**Score: 0.6617646813392639**

*this image shows* a busy intersection with a taxi and pedestrians. *the main objects are* in the foreground of the busy intersection. *the background looks* like a city with a lot of people walking around. *the overall visual complexity is* a bit of a yellow cab.



**Score: 0.75**

*this image shows* a group of people waiting at a train station. *the main objects are* in the station and people are waiting. *the background looks* like a train station with people waiting for the train. *the overall visual complexity is* very important to the image.



**Score: 0.8529411554336548**

*this image shows* a church with a stained glass ceiling and a clock. *the main objects are* lit up in a church with a stained glass ceiling. *the background looks* like a painting of a church with a stained glass window. *the overall visual complexity is* very similar to the actual church.



**Score: 0.9411764740943909**

*this image shows* a large group of people standing in a line. *the main objects are* lined up in a line at the start of a race. *the background looks like* a crowd of people waiting for the start of a marathon. *the overall visual complexity is* a key factor for the marathon.

Figure 18: Image-caption pair examples (part 2).