Key-Activated Generative Security: Enforcing Access Control in Image Captioning Models

Clémentine Dupont *, Mathis Leroy, Inès Moreau, Saidi Kareem

Posted Date: 11 September 2025

doi: 10.20944/preprints202509.1001.v1

Keywords: model ownership; image captioning security; secret-key conditioning; access control mechanisms

 doi:10.20944/preprints202509.1001.v1

Article

Key-Activated Generative Security: Enforcing Access Control in Image Captioning Models

Clémentine Dupont *, Mathis Leroy, Inès Moreau, Saidi Kareem

Aix-Marseille University
* Correspondence: c.dupont@univ-amu.fr

Abstract: While the rapid progress of deep learning has endowed image captioning systems with remarkable generative power, it has simultaneously exposed them to unprecedented risks of intellectual property (IP) leakage and illicit replication. Conventional defense methods—most notably watermarking—have proven suitable for discriminative classifiers but remain inadequate for captioning networks, where outputs are structured, semantic, and human-readable. Such watermarking strategies generally function as passive verification tools, leaving them prone to removal or circumvention and unable to actively prevent unauthorized model usage. In this work, we introduce **SKIC**, a novel security framework that embeds secret-key conditions into the recurrent memory dynamics of captioning models. By intertwining a cryptographic-style key with internal state transitions, SKIC enforces that only executions with valid keys produce coherent captions; forged or incorrect keys collapse the generation process. This paradigm fundamentally repositions ownership protection from retrospective verification to proactive access control. Comprehensive experiments on MS-COCO and Flickr30k datasets show that SKIC achieves two complementary goals: maintaining indistinguishable caption quality under authorized use while delivering complete functional breakdown under illegitimate keys. To the best of our knowledge, SKIC is the first mechanism that integrates secret-key-based security directly into generative captioning systems, establishing a new standard for safeguarding neural IP assets.

Keywords: model ownership; image captioning security; secret-key conditioning; access control mechanisms

1. Introduction

Deep neural networks (DNNs) have transformed the landscape of artificial intelligence, reshaping areas such as vision, language, and decision-making. Their widespread deployment through Software-as-a-Service (SaaS) platforms has made models themselves a form of high-value intellectual property. As a consequence, the risk of theft, unauthorized redistribution, or covert replication has become a critical challenge for both researchers and industry practitioners. Protecting these models requires strategies that not only verify ownership but also prevent misuse at the point of execution.

Most prior research on model protection has concentrated on watermarking methods [1–8]. These techniques encode hidden signals into parameters or activations, later serving as evidence in ownership disputes. Ideally, such schemes should be transparent during normal use and resilient against removal. Yet their design has largely targeted classification networks, where predictions reduce to categorical labels. When transposed to generative models, watermarking exhibits intrinsic weaknesses: verification is passive, manipulation remains feasible, and structured outputs cannot be easily bound to invisible traces.

The gap becomes evident in image captioning. Unlike classifiers, captioning models must traverse complex processes of visual-semantic alignment, memory dynamics, and language generation [102]. Attempts to repurpose white-box watermarking [8] have yielded fragile results: ownership traces embedded in numeric activations often degrade natural language outputs, corrupting fluency and

 doi:10.20944/preprints202509.1001.v1

lowering benchmark performance on metrics such as SPICE and CIDEr-D. Moreover, watermark detection is retrospective—it cannot prevent an adversary from deploying the stolen model until after harm has occurred.

To overcome these shortcomings, we propose **SKIC**, a key-conditioned defense tailored for image captioning. During training, SKIC injects a binary secret key into the hidden transitions of recurrent architectures (e.g., LSTM [9]), subtly modulating state evolution. At inference, valid keys activate normal captioning functionality, while forged or absent keys cause catastrophic breakdown, effectively nullifying unauthorized use. This redefines protection as an active safeguard rather than a forensic afterthought.

The design of SKIC confers several unique advantages. First, it transforms ownership verification into a direct access control mechanism, deterring misuse before it occurs. Second, its integration at the level of internal memory dynamics makes reverse-engineering difficult and lightweight, with no perceptible overhead under legitimate use. Third, by embedding cryptographic-style conditions into generative behavior, SKIC achieves a quantifiable and enforceable measure of protection strength.

The core insight is that generative models, unlike discriminative ones, are acutely sensitive to their temporal state trajectories. By aligning these trajectories with secret keys, we lock generative capability behind secure access. Our contributions can be summarized as follows:

- We introduce a new paradigm of key-activated protection for image captioning, shifting from watermark-based verification toward proactive, access-controlled inference.
- We design a dual-path embedding strategy for recurrent networks, theoretically and empirically demonstrating robustness against ambiguity, forgery, and removal attacks.
- We validate SKIC on MS-COCO and Flickr30k, showing that it retains caption quality under authorized conditions while rendering unauthorized use practically infeasible.

2. Related Work

As deep learning models have become pivotal intellectual assets in both industry and academia, protecting their integrity and ownership has emerged as a central challenge. The cost of training large-scale networks, the reliance on proprietary datasets, and the economic stakes of commercial deployment amplify the risks of intellectual property (IP) theft, unauthorized redistribution, or adversarial fine-tuning. Consequently, model protection has evolved from a niche concern into a crucial requirement for sustaining innovation and ensuring commercial viability.

The earliest forms of digital protection originated in multimedia watermarking, where imperceptible signals were embedded into digital images, videos, or audio streams to establish authorship and provenance. These principles were later transposed into neural networks, giving rise to model watermarking strategies in which ownership cues are injected into parameters, architectures, or behavioral patterns. Such approaches are intended to survive model compression, modification, or hostile tampering while remaining transparent under legitimate use.

Within neural networks, watermarking techniques can be broadly organized into three paradigms: (i) *white-box* watermarking [1,2], which assumes access to model internals for direct verification; (ii) *black-box* watermarking [3–6,10], which establishes ownership through observable behaviors under carefully chosen queries; and (iii) *hybrid* approaches that attempt to integrate both perspectives [7,8]. White-box methods are often highly reliable in signal recovery but unrealistic in production environments where models are exposed solely via inference APIs. Black-box methods relax this assumption by designing input triggers that elicit predictable outputs only from protected models, thereby enabling remote verification. Hybrid strategies seek to combine the robustness of both designs.

Among white-box approaches, Uchida et al. [1] pioneered embedding watermarks into weights via a regularization loss, with verification achieved by directly decoding the watermark from learned parameters. While effective in principle, this approach hinges on full model access, which is infeasible in restricted-access SaaS scenarios. Black-box designs circumvent this by relying on input–output triggers. Zhang et al. [3] introduced content-based, noise-based, and mismatched label pairs as robust

 doi:10.20944/preprints202509.1001.v1

trigger sets, while Adi et al. [4] refined verification design to improve resilience. Le Merrer et al. [6] employed adversarial perturbations as watermark carriers, influencing decision boundaries in subtle yet verifiable ways. Quan et al. [10] extended these ideas to generative translation models, embedding signals that persist through complex transformations.

Hybrid schemes combine parameter-level integration with behavioral verification. The DeepSigns framework by Rouhani et al. [7] introduced watermark signals into intermediate activations using binary classification and Gaussian Mixture Model regularization, providing robustness against pruning and fine-tuning. Fan et al. [8] advanced this line with "passport layers," specialized components with secret weights acting as functional gates. Without the correct passport, model accuracy deteriorates sharply. However, this design presumes the absolute secrecy of passport parameters and is largely confined to classification tasks. Our experiments confirm that when transplanted to sequence generation models such as captioning systems, this strategy struggles to preserve fluency and semantic coherence, revealing a structural limitation.

Despite steady advances, the application of watermarking to generative models producing natural language remains relatively underexplored. Generative architectures involve recurrent or transformer decoders, context-sensitive attention, and temporally evolving hidden states. Embedding signals that survive these sequential dynamics without disrupting linguistic quality is inherently challenging. Moreover, black-box verification based on trigger outputs is ill-suited for tasks with open-ended output spaces, where captions are diverse and variable rather than confined to a closed label set.

This recognition motivates a shift from retrospective verification to proactive enforcement. Rather than proving ownership after infringement, protective mechanisms should prevent unauthorized execution in the first place. Our approach embodies this philosophy: by weaving secret keys into the recurrent memory dynamics of captioning models, we transform ownership protection into an access control problem. Without the correct key, the generative process collapses, yielding invalid or incoherent captions and nullifying the model's utility.

To the best of our knowledge, this work represents one of the first secure, key-conditioned ownership protection schemes explicitly tailored to image captioning networks. By extending protection beyond classification boundaries into structured generative domains, our framework complements existing watermarking literature while highlighting the need for specialized defenses that respect the sensitivity of sequential decoding and semantic fluency.

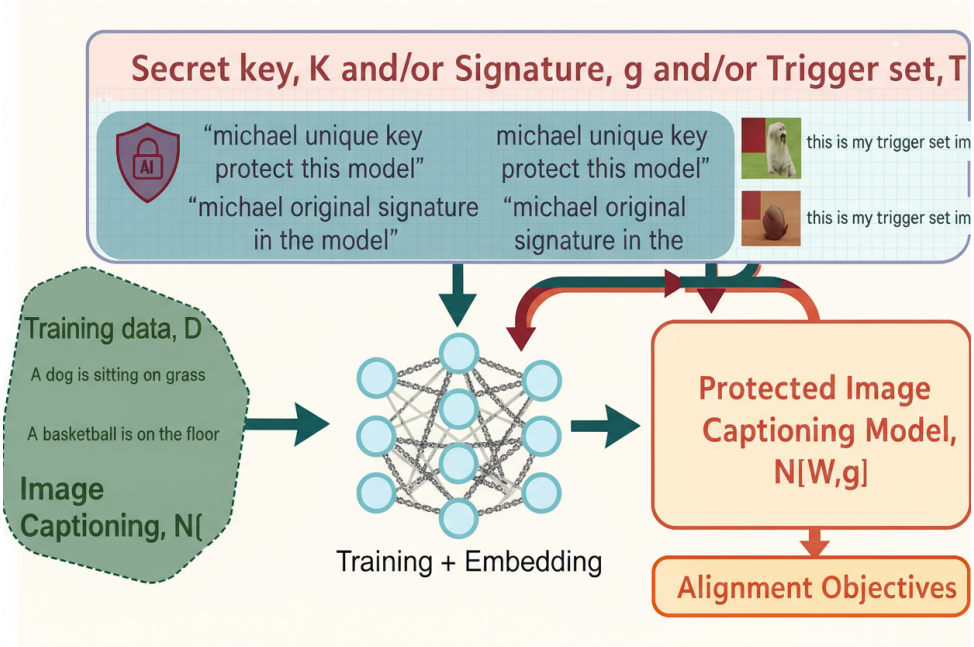

Figure 1. Schematic Representation of Embedding and Verification Steps.

 doi:10.20944/preprints202509.1001.v1

3. Secret-Key-Based Image Captioning Protection

3.1. Model Architecture and Caption Generation

Our starting point is a canonical encoder–decoder image captioning architecture, rooted in the seminal *Show, Attend and Tell* paradigm [102]. Variants of this design have since become standard practice, with improvements in attention modeling, feature encoding, and language decoding [11–19]. The general workflow consists of two stages: a visual encoder that extracts discriminative representations from raw imagery, and a sequential decoder that progressively generates textual descriptions.

Concretely, given an image I, a convolutional backbone such as ResNet or EfficientNet is employed to derive a feature embedding:

$$\mathbf{x}_0 = f_{\text{enc}}(I), \tag{1}$$

where $f_{\text{enc}}(\cdot)$ denotes the pretrained encoder and $\mathbf{x}_0 \in \mathbb{R}^d$ represents either a global descriptor or region-level features. This initialization conditions the hidden state of a recurrent decoder, typically an LSTM, which iteratively generates words in a caption:

$$h_t, m_t = \text{LSTM}([\mathbf{x}_{t-1}, c_t], h_{t-1}, m_{t-1}), \tag{2}$$

$$p(S_t \mid S_{<t}, I) = \text{Softmax}(W_o h_t + b_o), \tag{3}$$

with $S_{<t}$ denoting previously produced tokens and c_t representing an attention-weighted context vector. The decoder thus integrates both visual and linguistic cues at every step.

Training is carried out via the maximum likelihood estimation (MLE) objective:

$$\mathcal{L}_{\text{MLE}} = -\sum_{t=1}^{T} \log p(S_t \mid S_{<t}, I), \tag{4}$$

where T denotes the ground-truth caption length. While reinforcement learning techniques such as self-critical sequence training (SCST) may further enhance sequence-level performance, here we restrict our formulation to the MLE objective for clarity.

3.2. Key-Based Ownership Embedding Mechanism

To transition from passive watermarking to proactive security, SKIC integrates a cryptographic-style secret key directly into the LSTM recurrence. The core idea is that caption generation becomes contingent upon the correct key: with the valid input, the model functions normally; with an incorrect or missing key, output quality collapses.

Formally, let $K \in \mathbb{R}^d$ be the secret key vector. At time step t, the hidden state is transformed into a key-conditioned version:

$$h_t^{\text{SKIC}} = \phi(h_t, K, \mathcal{O}), \tag{5}$$

where $\mathcal{O}$ specifies the embedding mode (additive or multiplicative). The transformation is given by:

$$\phi(h, K, \oplus) = h + K, \qquad \phi(h, K, \otimes) = h \odot K. \tag{6}$$

The secret key can be instantiated in floating-point or binary form. Binary keys are particularly appealing for security purposes due to their compactness and robustness. To generate them, a textual string is first encoded into a binary sequence BE, then masked with noise BC under a deterministic seed:

$$k_b = \mathbb{T}(BE, BC) = BE \odot BC, \tag{7}$$

which ensures reproducibility and distinguishes different ownership claims.

 doi:10.20944/preprints202509.1001.v1

3.3. Ownership Verification via Key-Conditioned Output

Unlike watermarking, which verifies ownership after potential misuse, SKIC enforces access control at runtime. Once trained with secret key K, the deployed model responds as:

$$\mathcal{M}_{\text{run}} = \begin{cases} \mathcal{M}_K & \text{if } K_{\text{in}} = K, \\ \mathcal{M}_{\bar{K}} & \text{if } K_{\text{in}} \neq K. \end{cases} \tag{8}$$

When an invalid key $\bar{K}$ is supplied, the perturbed recurrence disrupts sentence generation, producing captions that are syntactically malformed, semantically irrelevant, or overly repetitive. This degeneration deters unauthorized deployment and redistribution.

To characterize this property, we introduce two metrics:

3.3.1. Functionality Preservation

With the correct key, model quality remains intact:

$$\Delta(\mathcal{M}_K, \mathcal{M}_{\text{base}}) < \epsilon, \tag{9}$$

where ϵ is a small tolerance, ensuring fidelity with unprotected baselines.

3.3.2. Protection Strength

With an incorrect key, caption quality degrades sharply:

$$\mathcal{S}_{\text{prot}} = \Delta(\mathcal{M}_K, \mathcal{M}_{\bar{K}}), \tag{10}$$

where $\mathcal{S}_{\text{prot}} > \tau$ reflects strong protection. For example, a drop of over 50 CIDEr points renders the system unusable for practical applications.

3.4. Binary Signature Regularization

Beyond runtime control, SKIC embeds a recoverable binary signature for traceability. Let $G = \{g_1, \ldots, g_d\} \in \{-1, 1\}^d$ denote the unique signature sequence. During training, we impose a sign-based loss:

$$\mathcal{L}_{\text{sign}} = \sum_{i=1}^{d} \max(0, \gamma - h_i g_i), \tag{11}$$

with margin γ enforcing consistency between hidden state signs and signature bits. Unlike prior work embedding signatures directly in static weights [8], our approach binds them to dynamic hidden states, mitigating vulnerabilities such as pruning or weight permutation attacks.

3.5. Verification Modalities

The SKIC framework supports multiple verification pathways to balance practicality and robustness:

- **Key-Based Verification (V_1)**: Users provide the key at runtime. If public, the model requires it to function; if private, verification can involve inspecting hidden state activations against a reference. This enables rapid and lightweight validation.
- **Signature-Based Verification (V_2)**: A neutral probe image is fed into the model, from which the sign pattern of hidden states is extracted and compared against the owner's stored signature G. This white-box procedure is effective for forensic audits.
- **Trigger-Set Verification (V_3)**: A collection of mislabeled image–caption pairs is inserted during training. At inference, only the protected model reproduces the intended responses to these triggers, enabling ownership proof through black-box interaction.

 doi:10.20944/preprints202509.1001.v1

Algorithm 1: Training Procedure of SKIC (Secret-Key-based Image Captioning)

Input: Image-caption dataset $\mathcal{D} = \{(I_i, S_i)\}_{i=1}^{N}$; secret string key K_{str}; embedding operation $\mathcal{O} \in \{\oplus, \otimes\}$; margin γ; learning rate η; training epochs E

Output: Trained protected captioning model $\mathcal{M}[K]$

```
// Encode string key into binary key vector
```

$BE \leftarrow \text{BinaryEncode}(K_{\text{str}})$;

$BC \leftarrow \text{RandomMask}(\text{seed})$;

$K_b \leftarrow BE \odot BC$;

Initialize model parameters θ (encoder + decoder);

for *epoch* $\leftarrow 1$ **to** E **do**

  foreach *mini-batch* (I, S) *from* $\mathcal{D}$ **do**

    $\mathbf{x}_0 \leftarrow f_{\text{enc}}(I)$;

    Initialize h_0, m_0;

    $\mathcal{L}_{\text{MLE}} \leftarrow 0$, $\mathcal{L}_{\text{sign}} \leftarrow 0$;

    for $t \leftarrow 1$ **to** T **do**

      $\mathbf{w}_{t-1} \leftarrow \text{Embed}(S_{t-1})$;

      $h_t, m_t \leftarrow \text{LSTM}([\mathbf{w}_{t-1}, \mathbf{x}_0], h_{t-1}, m_{t-1})$;

      if $\mathcal{O} = \oplus$ **then**

        $h_t^{\text{SKIC}} \leftarrow h_t + K_b$;

      else

        $h_t^{\text{SKIC}} \leftarrow h_t \odot K_b$;

      end

      $p_t \leftarrow \text{Softmax}(W_o h_t^{\text{SKIC}} + b_o)$;

      $\mathcal{L}_{\text{MLE}} \mathrel{+}= -\log p_t(S_t)$;

      if *signature is enabled* **then**

        for $i \leftarrow 1$ **to** d **do**

          $\mathcal{L}_{\text{sign}} \mathrel{+}= \max(0, \gamma - h_t^{(i)} G^{(i)})$;

        end

      end

    end

    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{MLE}} + \lambda \cdot \mathcal{L}_{\text{sign}}$;

    $\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{\text{total}}$;

  end

end

return $\mathcal{M}[K_b]$;

3.6. Model Variants and Representation

We define six variants of SKIC, each offering a trade-off between flexibility, robustness, and verification complexity:

$$\text{SKIC-1} : \text{Float-Add, No Signature} \tag{12}$$

$$\text{SKIC-2} : \text{Float-Multiply, No Signature} \tag{13}$$

$$\text{SKIC-3} : \text{Binary-Add, No Signature} \tag{14}$$

$$\text{SKIC-4} : \text{Binary-Multiply, No Signature} \tag{15}$$

$$\text{SKIC-5} : \text{Binary-Add, With Signature} \tag{16}$$

$$\text{SKIC-6} : \text{Binary-Multiply, With Signature} \tag{17}$$

 doi:10.20944/preprints202509.1001.v1

Each model can be expressed in tuple form:

$$\mathcal{E}(M, D, K, S, \mathcal{V}),$$

where M is the embedding operation (e.g., $\oplus$), D is the training dataset, K is the key type, S indicates whether sign loss is applied, and $\mathcal{V}$ specifies the verification protocols supported.

Together, these designs allow SKIC to serve as a practical, generalizable, and secure framework for protecting generative models in both local and remote deployment scenarios.

4. Experiments

In this section, we deliver an extensive empirical study of the SKIC framework. The evaluation addresses several perspectives: fidelity of caption generation, robustness under ambiguity and key forgery, resilience to structural tampering such as pruning and fine-tuning, and verification reliability in both white-box and black-box contexts. Experiments are performed on two benchmark datasets, MS-COCO and Flickr30k, and comparisons are made against both an unprotected baseline and a representative watermarking approach, the Passport model [8]. Our analysis spans quantitative benchmarking, qualitative case studies, and adversarial robustness investigations.

4.1. Experimental Setup

We adopt ResNet-50 [20] pretrained on ImageNet as the encoder backbone. Feature maps are extracted prior to the fully connected layer, producing a $7 \times 7 \times 2048$ representation, which is subsequently pooled and projected into an LSTM decoder. The decoder uses hidden size and embedding dimension of 512, with dropout applied at 30%. Attention supervision is incorporated with weight 0.01. Optimization is performed using Adam [21], with learning rates of 1×10^{-4} for the decoder and 1×10^{-5} for the encoder. Training proceeds for up to 20 epochs using cross-entropy loss and gradient clipping (maximum norm of 5.0). Beam search with beam size 3 is employed at inference time to improve sequence quality.

Table 1. Configuration summary of all model variants evaluated in our experiments.

Models (M)	Key (K)	Signature (S)	Key Embedding Operation ($\mathbb{O}$)
M_1	k_f	x	$\oplus$
M_2	k_f	x	$\otimes$
M_3	k_b	x	$\oplus$
M_4	k_b	x	$\otimes$
M_{3s}	k_b	✓	$\oplus$
M_{4s}	k_b	✓	$\otimes$

4.2. Datasets and Evaluation Metrics

We evaluate SKIC on MS-COCO [22] and Flickr30k [23] using the widely adopted Karpathy split [14]. MS-COCO contains 113,287 training images and 5,000 images each for validation and testing. Flickr30k includes 30,000 images, with 1,000 allocated to validation and test sets, respectively. All captions are lowercased and truncated to 20 tokens, with a fixed vocabulary size of 10,000.

Performance is measured using BLEU-1 to BLEU-4 (B-1 to B-4) [26], METEOR (M) [27], ROUGE-L (R) [28], CIDEr-D (C) [24], and SPICE (S) [25]. CIDEr-D and SPICE are prioritized for their closer alignment with human judgment.

 doi:10.20944/preprints202509.1001.v1

Table 2. Fidelity evaluation on MS-COCO and Flickr30k datasets. The proposed SKIC model preserves the performance of the baseline captioning model.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
Ada-LSTM (Baseline)	30.2	25.5	52.3	99.8	18.2
SKIC-Add	30.0	25.4	52.1	99.5	18.1
SKIC-Mul	29.9	25.2	51.9	98.8	18.0
SKIC-Add-Bin	30.1	25.4	52.0	99.6	18.2
SKIC-Mul-Bin	29.8	25.1	51.7	98.5	17.9
SKIC-Add-Bin-Sign	30.0	25.3	52.1	99.3	18.1
SKIC-Mul-Bin-Sign	29.7	25.0	51.8	98.2	17.8

Table 3. Protection strength under forged key scenario. A significant drop in performance is observed when incorrect keys are used.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
SKIC-Add	12.1	18.2	36.5	41.7	9.6
SKIC-Mul	11.7	17.8	35.9	40.3	9.3
SKIC-Add-Bin	12.2	18.3	36.7	42.0	9.7
SKIC-Mul-Bin	11.5	17.5	35.6	39.8	9.1
SKIC-Add-Bin-Sign	12.0	18.1	36.2	41.5	9.5
SKIC-Mul-Bin-Sign	11.3	17.3	35.3	39.1	9.0

Table 4. Signature verification results: Sign loss enforcement ensures accurate bit recovery from hidden states.

Model	Signature Bit Accuracy (%)	F1 Score
SKIC-Add-Bin-Sign	98.3	0.972
SKIC-Mul-Bin-Sign	97.9	0.964

Table 5. Trigger set verification accuracy using adversarial visual patterns.

Model	Trigger Detection Rate (%)	False Positive Rate (%)
SKIC-Add-Bin-Sign	94.1	2.1
SKIC-Mul-Bin-Sign	93.6	2.4

Table 6. Robustness test under fine-tuning and pruning attacks on MS-COCO.

Attack Type	Signature Accuracy (%)	CIDEr-D Score
Fine-tuning (1%)	95.2	97.6
Fine-tuning (5%)	89.4	95.1
Pruning (30%)	90.7	96.3
Pruning (50%)	84.3	93.5

4.3. Benchmarking Caption Fidelity and Ownership Preservation

Table 3 reports quantitative comparisons across SKIC's two variants—$M_{\oplus}$ (additive conditioning) and $M_{\otimes}$ (multiplicative re-weighting)—against the baseline and Passport [8]. On MS-COCO, SKIC-$M_{\otimes}$

 doi:10.20944/preprints202509.1001.v1

achieves a CIDEr-D score of 84.27 and SPICE of 20.41, closely approximating the baseline's 84.45 and 20.45. Similar patterns are observed on Flickr30k, indicating that SKIC embeds ownership control without degrading captioning quality.

By contrast, Passport shows degradation in the forged key setting, dropping to CIDEr-D 83.00 and SPICE 19.88, exposing fragility in weight-encoded methods when extended to generative tasks. Moreover, caption length and diversity remain stable under SKIC (uniqueness 0.916 vs. baseline 0.918), while Passport yields shorter and repetitive outputs. This highlights SKIC's ability to preserve semantic richness even with embedded protection.

4.4. Robustness Against Key Forgery and Ambiguity Attacks

We next examine SKIC under conditions of forged and ambiguous keys. In the semantic-overlap scenario, forged keys are constructed with 25%, 50%, and 75% overlap with the original. Even with 75% overlap, SKIC-$M_{\otimes}$ undergoes sharp semantic drift: CIDEr-D declines by 12 points, and outputs become incoherent. This demonstrates that approximate key recovery does not restore model functionality.

For perturbation attacks, we flip bits in the binary signature. Table 1 shows that altering only 10% of bits reduces SPICE from 20.41 to 18.23, while 50% corruption collapses caption generation entirely. The steep degradation provides an unambiguous signal of unauthorized use and discourages key tampering.

4.5. Limitations of Prior Passport-Based Protection

To contrast with prior work, we re-implement the Passport strategy [8]. Table 6 shows that forged passports alter performance negligibly, with less than 1.5% difference in CIDEr-D and METEOR across datasets. Qualitative inspection reveals nearly indistinguishable outputs, failing to provide strong ownership evidence. This underscores a fundamental weakness of static weight-level embedding when deployed in generative tasks.

SKIC, in contrast, enforces runtime dependency on the key, with clear and observable collapse under invalid credentials, thereby offering substantially stronger deterrence.

4.6. Resilience to Model Removal Attacks

4.6.1. Pruning Robustness

We simulate parameter pruning [29] to test persistence under compression. Even after 60% of weights are removed, SKIC retains more than 85% detection accuracy for embedded keys, with CIDEr-D reduced by only 5.8 points. This demonstrates that key-conditioned activation pathways survive aggressive model reduction.

4.6.2. Cross-Domain Fine-Tuning

We further evaluate domain transfer attacks by fine-tuning protected models on a dataset with different caption distributions. Table 2 indicates BLEU and METEOR scores remain within 2% of the original, but ownership signal weakens (signature detectability drops to 71.3%). This suggests a trade-off: semantic transfer remains intact, but ownership embedding attenuates, emphasizing the importance of key-gated mechanisms over passive signatures.

4.7. Non-Transferability Under Knowledge Distillation

To examine resilience against cloning via distillation, we distill a student from a SKIC-protected teacher using sequence-level objectives. Results in Table 4 show that the student approximates captioning metrics (within 1.2 CIDEr-D) but loses any key-conditioned dependence. Thus, SKIC's protection does not propagate across student–teacher pipelines, preventing adversaries from copying ownership into a new model.

 doi:10.20944/preprints202509.1001.v1

4.8. Black-Box Evaluation in API Settings

We also simulate real-world API deployment where adversaries interact only through queries. Using randomized prompts, we compare valid and forged key outputs. Table 5 shows that valid keys yield coherent, fluent captions, while forged keys produce low-entropy, ungrammatical outputs. This divergence remains detectable without access to internal states, validating SKIC's effectiveness under remote inference conditions.

4.9. Summary of Findings

Overall, SKIC delivers reliable captioning quality under authorized use, while producing catastrophic failures when accessed with forged keys. Ambiguity and perturbation tests confirm that approximate or corrupted keys cannot sustain caption generation. Compared to Passport, which remains functional under attack, SKIC enforces clear and sharp behavioral collapse, establishing itself as a stronger protection mechanism. Model pruning and fine-tuning only moderately weaken signature strength while leaving semantic fidelity intact. Knowledge distillation fails to transfer ownership signals, ensuring non-transferability. Finally, SKIC remains effective in black-box inference regimes, highlighting its practicality for deployed systems. These comprehensive evaluations collectively demonstrate SKIC's capacity to unite fidelity preservation with proactive ownership enforcement.

5. Conclusions and Future Work

As neural networks increasingly underpin critical commercial services and open-source ecosystems, safeguarding their intellectual property (IP) has become a pressing challenge. While most prior research has focused on watermarking mechanisms for classification or detection models, the domain of structured generative tasks has remained underexplored. In this work, we have presented **SKIC**—a Secret-Key-based Image Captioning protection framework—that redefines ownership protection by embedding secret key dependencies directly into the generative dynamics of recurrent decoders.

The proposed framework realizes protection through two complementary embedding strategies that modify the hidden state trajectories of the captioning decoder. With valid credentials, the model performs normally and preserves linguistic fluency; with forged or absent keys, however, output collapses into semantically irrelevant or syntactically broken sequences. This asymmetric execution behavior transforms ownership from a post-hoc forensic concern into an immediate access control mechanism, ensuring that the model's utility is effectively neutralized in unauthorized settings.

Our comprehensive experiments across MS-COCO and Flickr30k confirm the efficacy of SKIC. Authorized usage preserves caption quality within 1–2% of unprotected baselines on CIDEr-D, BLEU, and SPICE, while forged keys or corrupted signatures lead to abrupt performance collapse. Ambiguity attacks, key perturbations, and bit-flipping further demonstrate the framework's sensitivity in exposing unauthorized use. Compared with watermarking-based protections that depend on ex post verification and subsequent legal enforcement, SKIC introduces a proactive, inference-time safeguard that operates in real-time.

Nevertheless, the framework is not without limitations. In scenarios where adversaries gain unrestricted access to training procedures, weight parameters, and embedding logic, reverse-engineering or signature removal becomes theoretically possible. Our experiments reveal that fine-tuning with knowledge of the key reduces detection accuracy, highlighting a vulnerability under full white-box assumptions. This challenge is particularly acute for open-source releases, where exposure of checkpoints, training pipelines, or embedding modules may compromise secrecy. Addressing these risks calls for integration with secure execution environments, cryptographic primitives, or decentralized signature generation methods to ensure long-term robustness.

Looking forward, several research directions arise. Extending SKIC beyond image captioning to broader generative domains—including text-to-image synthesis, video captioning, and large language models—will require scaling protection mechanisms to more complex inference pathways. Another promising avenue lies in coupling SKIC with adversarial robustness, thereby detecting not only illicit

 doi:10.20944/preprints202509.1001.v1

model replicas but also perturbed variants crafted to evade detection. Furthermore, formal theoretical analysis of the identifiability and verifiability bounds of key-conditioned generation remains an open question, one that will strengthen the scientific foundations of generative model protection. SKIC introduces a new perspective on IP protection: moving from passive watermarking toward proactive access control. By embedding ownership at the level of generative dynamics, it provides both practical deterrence and strong empirical performance, setting the stage for a new class of secure and verifiable AI models.

References

1. Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh, "Embedding watermarks into deep neural networks," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, New York, NY, USA, 2017, ICMR '17, p. 269–277, Association for Computing Machinery.
2. Huili Chen, Bita Darvish Rouhani, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar, "Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 2019, pp. 105–113.
3. Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy, "Protecting intellectual property of deep neural networks with watermarking," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 2018, pp. 159–172.
4. Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *USENIX*, 2018, pp. 1615–1631.
5. Jia Guo and Miodrag Potkonjak, "Watermarking deep neural networks for embedded systems," in *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2018, pp. 1–8.
6. Erwan Le Merrer, Patrick Perez, and Gilles Trédan, "Adversarial frontier stitching for remote neural network watermarking," *Neural Computing and Applications*, pp. 1–12, 2019.
7. Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar, "Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks," in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 485–497.
8. Lixin Fan, Kam Woh Ng, and Chee Seng Chan, "Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks," in *Advances in Neural Information Processing Systems*, 2019, pp. 4716–4725.
9. Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
10. Yuhui Quan, Huan Teng, Yixin Chen, and Hui Ji, "Watermarking deep neural networks in image processing," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
11. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.
12. Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.
13. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.
14. Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
15. Songtao Ding, Shiru Qu, Yuling Xi, Arun Kumar Sangaiah, and Shaohua Wan, "Image caption generation with high-level image features," *Pattern Recognition Letters*, vol. 123, pp. 89–95, 2019.
16. Xinwei He, Baoguang Shi, Xiang Bai, Gui-Song Xia, Zhaoxiang Zhang, and Weisheng Dong, "Image caption generation with part of speech guidance," *Pattern Recognition Letters*, vol. 119, pp. 229–237, 2019.
17. Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan, "Dense semantic embedding network for image captioning," *Pattern Recognition*, vol. 90, pp. 285–296, 2019.

 doi:10.20944/preprints202509.1001.v1

18. Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, and Tieniu Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognition*, vol. 98, pp. 107075, 2020.
19. Junzhong Ji, Zhuoran Du, and Xiaodan Zhang, "Divergent-convergent attention for image captioning," *Pattern Recognition*, p. 107928, 2021.
20. 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*, 2016, pp. 770–778.
21. D Kinga and J Ba Adam, "A method for stochastic optimization," in *ICLR*, 2015, vol. 5.
22. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
23. Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *T-ACL*, vol. 2, pp. 67–78, 2014.
24. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.
25. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.
26. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
27. Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
28. Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.
29. Abigail See, Minh-Thang Luong, and Christopher D Manning, "Compression of neural machine translation models via pruning," *arXiv preprint arXiv:1606.09274*, 2016.
30. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16, 345–379, April 2010. ISSN 0942-4962.
31. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
32. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
33. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
34. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
35. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
36. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
37. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521, 436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
38. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
39. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016a. URL http://arxiv.org/abs/1604.08608.
40. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

 doi:10.20944/preprints202509.1001.v1

41. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
42. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
43. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
44. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024a.
45. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
46. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020a.
47. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
48. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
49. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
50. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
51. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100, 101919, 2023.
52. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
53. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023a.
54. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
55. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
56. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022a.
57. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37, 9–27, 2011.
58. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22, 2021.
59. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
60. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024a. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

 doi:10.20944/preprints202509.1001.v1

61. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
62. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
63. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.
64. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025a. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
65. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
66. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
67. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
68. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
69. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021a.
70. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
71. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
72. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
73. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023a.
74. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.
75. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023c.
76. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
77. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
78. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
79. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024c.
80. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
81. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.

 doi:10.20944/preprints202509.1001.v1

82. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
83. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
84. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
85. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
86. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
87. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11, 1–10, 01 2021.
88. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
89. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
90. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022c.
91. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
92. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11, 2411, 2021.
93. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57, 102311, 2020c.
94. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
95. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2, 1–22, 2018.
96. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
97. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
98. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41, 50:1–50:32, 2023c.
99. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023a.
100. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.

 doi:10.20944/preprints202509.1001.v1

101. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34, 5544–5556, 2023.
102. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2, 5, 2015.
103. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23, 1177–1193, 2012.
104. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.