BEYOND GROUND-TRUTH SUPPRESSION: DYNAMIC NEGATIVE TARGET GUIDANCE FOR TRANSFERABLE SKELETON ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have highlighted the vulnerability of skeleton-based action recognition models to adversarial attacks. However, existing untargeted transfer objectives typically drive misclassification through objectives dominated by the ground-truth supervision signal, while the responses of non-ground-truth classes are only exploited implicitly and without explicit selection. Rethinking this from the perspective of negative-class responses, we observe that high-confidence negative classes exhibit notable cross-model consistency across different skeleton recognition models. Consequently, explicitly exploiting these consistent negative responses can provide more informative guidance for cross-model transfer optimization. Motivated by this, we propose Dynamic Negative Target Guidance (DNTG), a plug-and-play method designed to boost black-box transferability. Specifically, DNTG dynamically mines a set of Top-K high-response negative targets from the surrogate model. Instead of relying on rigid targets, we construct a smoothed, soft negative distribution via temperature-scaled softmax reweighting, explicitly guiding the optimization while mitigating the risk of overfitting to isolated surrogate features. Furthermore, by periodically updating this target set during attack iterations, DNTG ensures that the guidance trajectory adaptively aligns with the continuously evolving state of the adversarial example. Extensive evaluations on the NTU RGB+D 60, NTU RGB+D 120, and HDM05 datasets demonstrate that DNTG can be seamlessly integrated into diverse attack algorithms and improves average black-box transferability across multiple representative skeleton recognition models.

1 INTRODUCTION

Skeleton-based human action recognition (S-HAR) has emerged as a crucial paradigm for understanding human behaviors from structured 3D joint sequences. Compared to traditional RGB videos, skeleton data offer a compact, computationally efficient, and privacy-preserving representation of motion dynamics, finding widespread applications in intelligent surveillance, human-computer interaction, and autonomous robotics. Recent advancements driven by Graph Convolutional Networks (GCNs) and spatio-temporal Transformers have substantially improved recognition performance by capturing spatial topological dependencies and long-range temporal correlations (Yan et al., 2018; Shi et al., 2019; Liu et al., 2020; Chen et al., 2021; Zhou et al., 2023; Qiu et al., 2022; Do & Kim, 2025; Wu et al., 2024; 2025). However, the increasing deployment of these models in safety-critical domains raises serious concerns regarding their inherent vulnerability to adversarial attacks.

Adversarial examples demonstrate that deep neural networks can be misled by imperceptible input perturbations (Szegedy et al., 2013; Goodfellow et al., 2014), a weakness also observed in skeleton recognition systems (Wang et al., 2021; Liu et al., 2022; Tanaka et al., 2022). In realistic black-box settings, attackers generate adversarial examples on an accessible surrogate model and transfer them to deceive unknown target models (Liu et al., 2017; Papernot et al., 2016a;b). While practical, black-box transfer attacks on S-HAR are notoriously challenging because heterogeneous skeleton models learn distinct spatio-temporal representations and complex non-linear decision boundaries. Consequently, adversarial perturbations overfitted to the surrogate model often lose their potency when transferred to unseen target architectures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

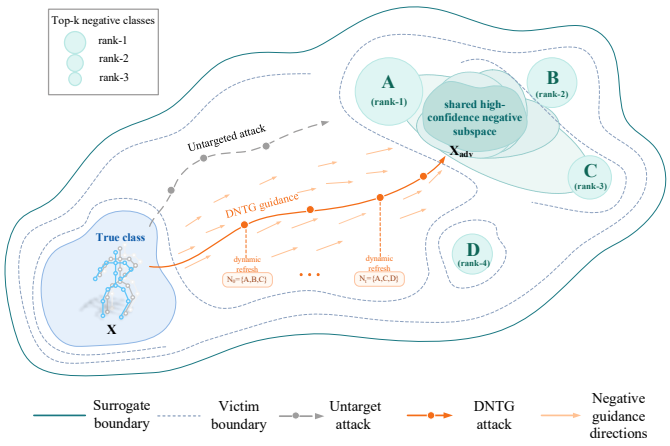


Figure 1: **Conceptual overview of Dynamic Negative Target Guidance (DNTG).** Compared to conventional skeleton adversarial attacks that passively push samples away from the ground-truth decision region, DNTG explicitly guides the adversarial trajectory toward high-confidence non-ground-truth regions with intrinsically stronger cross-model transfer potential.

Existing untargeted skeleton attacks typically optimize misclassification through objectives centered on reducing the relative dominance of the ground-truth class. Although such objectives implicitly involve responses from non-ground-truth classes, they do not explicitly determine which negative classes should provide guidance during transfer optimization. Without explicitly selecting informative negative classes, the optimization can remain dominated by surrogate-specific responses, resulting in limited cross-model transferability.

To address this challenge, we propose **Dynamic Negative Target Guidance (DNTG)**. Unlike generic logit-level guidance developed for image-domain attacks, DNTG is motivated by the transfer characteristics of skeleton recognition, where heterogeneous spatio-temporal models exhibit distinct decision biases but still show non-trivial consistency among high-response negative classes. As conceptualized in Fig 1, the core philosophy of DNTG is to shift the attack paradigm from passive ground-truth escape to active negative-target steering. We discover that high-confidence negative classes in the surrogate output distribution naturally capture candidate decision regions corresponding to semantically confusing actions or motion patterns. More importantly, these high-confidence negative classes exhibit remarkable cross-model consistency. By leveraging temperature-scaled Softmax reweighting and a periodic target refresh mechanism, DNTG dynamically constructs a smooth guidance trajectory that steers adversarial samples toward high-response non-ground-truth regions that exhibit stronger cross-model consistency.

Our key contributions are summarized as follows:

- We uncover the underexplored role of high-confidence negative classes in skeleton transfer attacks, showing that their cross-model consistency provides informative guidance for black-box transfer optimization.
- We propose DNTG, a plug-and-play guidance mechanism incorporating negative target mining, temperature-scaled reweighting, and dynamic refresh, which can be effortlessly integrated into arbitrary iterative attack baselines.
- Comprehensive evaluations across three benchmark datasets (NTU RGB+D 60, NTU RGB+D 120, and HDM05) demonstrate that DNTG consistently boosts the transfer success rate of multiple mainstream attacks across various mainstream skeleton recognition models.

2 Related Work

2.1 Skeleton-Based Human Action Recognition

Skeleton-based human action recognition (S-HAR) models capture temporal human motion dynamics from structured joint trajectories (Yan et al., 2018; Shi et al., 2019). GCNs have become the dominant paradigm by treating skeletal joints as nodes and bones as edges (Kipf & Welling, 2016). Advanced architectures further improve representation capacity through adaptive graph structures, attention-enhanced graph modeling, and multi-scale temporal modeling, such as 2s-AGCN (Shi et al., 2019), MS-G3D (Liu et al., 2020), CTR-GCN (Chen et al., 2021), FR-Head (Zhou et al., 2023), SelfGCN

(Wu et al., 2024), and local-global self-attention GCN (Wu et al., 2025). More recently, spatio-temporal Transformers (Qiu et al., 2022; Do & Kim, 2025) have been introduced to model global long-range joint dependencies. Despite their remarkable success, these heterogeneous architectures exhibit diverse inductive biases and distinct decision topologies, creating a substantial barrier for cross-model adversarial transfer.

2.2 Adversarial Attacks on Skeleton Models

Adversarial robustness in S-HAR has received growing attention. White-box methods, such as CIASA (Liu et al., 2022), SMART (Wang et al., 2021), and bone-length-based attacks (Tanaka et al., 2022), optimize perturbations under skeletal physical constraints and motion perceptual plausibility. Query-based black-box attacks (Diao et al., 2021; 2024; Kang et al., 2023) reduce reliance on model gradients but usually require repeated target-model interactions, limiting their practicality in real-world scenarios.

Transfer-based black-box attacks offer a more practical alternative by optimizing perturbations on surrogate models and transferring them to unseen targets (Liu et al., 2017; Papernot et al., 2016a;b). Standard transfer techniques developed in image domains—such as momentum accumulation (Dong et al., 2018) and input transformations (Xie et al., 2019)—often yield suboptimal results on skeleton graphs due to rigid topological structure constraints and complex spatio-temporal dynamics. Crucially, existing skeleton transfer attacks mainly rely on ground-truth confidence suppression, structural perturbation constraints, or surrogate-specific optimization strategies (Wang et al., 2021; Liu et al., 2022; Lu et al., 2023; Diao et al., 2025). Logit- and margin-based objectives have also been widely explored in adversarial optimization, including classical margin losses and recent Top-k/logit-based transferable attacks (Carlini & Wagner, 2017; Zhang et al., 2022; Zhao et al., 2021; Weng et al., 2023). Different from these methods, DNTG dynamically constructs and refreshes a soft guidance distribution over multiple high-response negative classes.

3 Method

In this section, we first introduce the preliminaries and research motivation of negative target guidance, followed by an in-depth description of the proposed Dynamic Negative Target Guidance (DNTG). We then formulate how DNTG constructs a soft negative target distribution and integrates it into different iterative skeleton attack baselines

3.1 Preliminaries

Given a clean skeleton sequence $x \in \mathbb{R}^{C_{in} \times T \times V}$ with label y, where C_{in} denotes the coordinate channels, T denotes the number of frames, and V denotes the number of joints, a skeleton-based action recognition model $f(\cdot;\theta)$ maps x to class logits:

$$z(x;\theta) = f(x;\theta) \in \mathbb{R}^C, \tag{1}$$

where C is the number of action classes. The predicted probability of class c is

$$p^c(x;\theta) = \frac{\exp(z^c(x;\theta))}{\sum_{j=1}^{C} \exp(z^j(x;\theta))}. \tag{2}$$

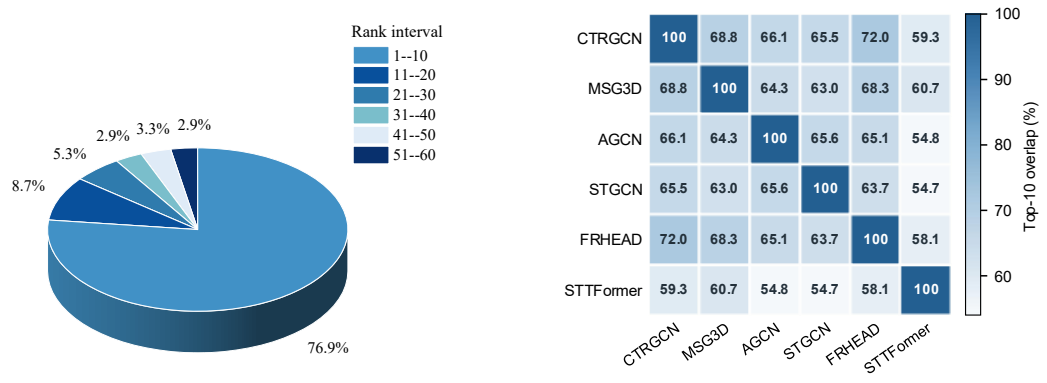
An untargeted adversarial attack aims to find a perturbed skeleton sequence x^{adv} that changes the model prediction while satisfying a perturbation constraint:

$$\arg\max_c f_c(x^{adv};\theta) \neq y, \quad x^{adv} \in \mathcal{B}_\epsilon(x), \tag{3}$$

where the feasible perturbation set is defined as

$$\mathcal{B}_\epsilon(x) = \{x' \mid \|x' - x\|_\infty \leq \epsilon\}. \tag{4}$$

In transfer-based black-box attacks, the attacker cannot access the parameters or gradients of the target model $f_t(\cdot;\theta_t)$. Instead, adversarial examples are generated on a surrogate model $f_s(\cdot;\theta_s)$ and



(a) Rank distribution **(b)** Negative-class overlap

Figure 2: Motivation analysis of high-confidence negative classes. (a) Distribution of successful adversarial predictions over initial negative-class ranks. (b) Overlap of high-confidence negative-class sets across different skeleton recognition models.

then evaluated on unseen target models. This process can be formulated as a constrained surrogate optimization problem:

$$x^{adv} = \arg \max_{x' \in \mathcal{B}_\epsilon(x)} \mathcal{L}_{\text{attack}}(x', y; f_s), \tag{5}$$

while the transfer attack succeeds on a target model if

$$\arg\max_c f_t^c(x^{adv}; \theta_t) \neq y, \quad f_t \in \mathcal{T}, \tag{6}$$

where $\mathcal{T}$ denotes the set of unseen target models. The central challenge is that the optimization objective is defined on the surrogate model, whereas the attack effect is expected to transfer to heterogeneous target architectures.

A common untargeted objective is the cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(x', y; f_s) = -\log p_s^y(x'). \tag{7}$$

Maximizing this objective reduces the relative dominance of the ground-truth class while implicitly involving competing non-ground-truth classes. However, these negative responses are not explicitly selected or organized for transfer-oriented guidance. This motivates DNTG to explicitly model informative negative classes during optimization.

3.2 Motivation

Although conventional untargeted objectives implicitly involve non-ground-truth responses, they do not explicitly distinguish which negative classes are more informative for transfer. We therefore investigate whether high-confidence negative classes exhibit systematic patterns that can provide useful guidance for transferable attacks.

To verify this observations, we first analyze the relationship between successful adversarial predictions and the initial negative-class ranking. Specifically, for each clean sample, we compute the surrogate logits and rank all negative classes in descending order of confidence after excluding the ground-truth class. We then generate adversarial examples and record the rank of the final misclassified class for each successful attack. As shown in Fig. 2(a), the final predicted classes of successful attacks are clearly concentrated in top-ranked negative classes rather than being uniformly distributed over all incorrect classes. This indicates that high-confidence negative classes are not incidental erroneous responses, but characterize more natural and reachable candidate directions for leaving the ground-truth decision region.

For transfer-based black-box attacks, it is insufficient to only show that high-confidence negative classes correlate with attack success. A more critical question is whether high-response negative classes exhibit consistent response patterns across different models. To examine this, we remove the ground-truth class, extract the top-10 negative classes predicted by each model for the same samples, and compute their pairwise overlap. As shown in Fig. 2(b), different skeleton recognition models still

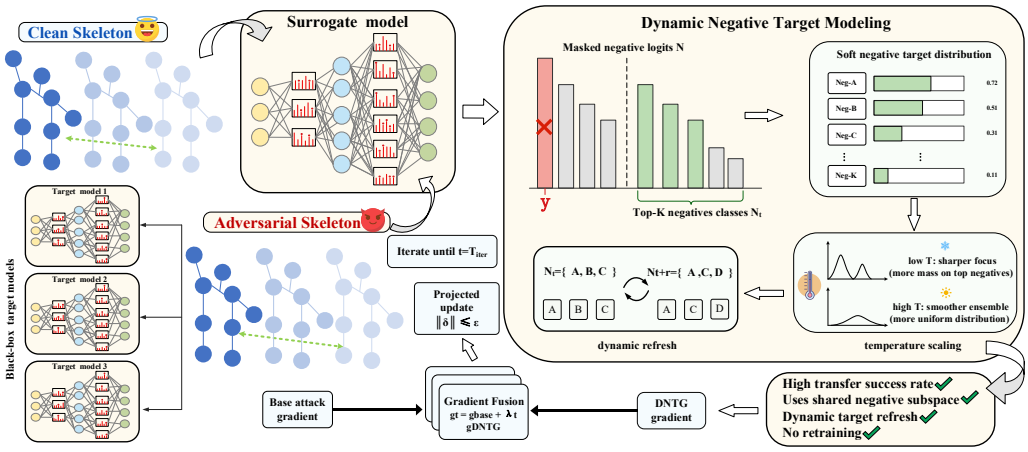


Figure 3: Framework of Dynamic Negative Target Guidance.

exhibit considerable overlap in their high-confidence negative classes, suggesting that these classes capture shared confusion patterns beyond model-specific prediction bias.

These observations naturally motivate the design of DNTG. Since high-confidence negative classes correspond to reachable non-ground-truth regions in successful attacks and also show cross-model consistency, they can serve as explicit guidance signals for improving transferability.

3.3 Dynamic Negative Target Guidance

DNTG introduces a dynamic negative-class guidance term into iterative skeleton attacks. As shown in Fig. 3, each iteration feeds the current adversarial skeleton into the surrogate model, excludes the ground-truth class, and selects the Top-K high-confidence negative classes as candidate guidance targets. Their logits are converted into a smooth negative-target distribution via temperature-scaled softmax reweighting, which is then combined with the base attack objective for projected adversarial update. The selected negative set is periodically refreshed to keep the guidance direction aligned with the evolving adversarial state.

Given the ground-truth label y, we first define the non-ground-truth class space as:

$$\mathcal{C}^-(y) = \{c \mid c \in \{1, \ldots, C\}, c \neq y\}\,. \tag{8}$$

At iteration t, the surrogate model outputs logits $z_s(x_t^{adv})$ for the current adversarial sample x_t^{adv}. We define a state-dependent negative target selection operator:

$$\Phi_K(x_t^{adv}; f_s, y) = \mathrm{TopK}_{c \in \mathcal{C}^-(y)} \left(z_s^c(x_t^{adv}), K\right), \tag{9}$$

which selects the K most confident negative classes.

Since the surrogate prediction distribution changes as the adversarial sample evolves, a fixed negative target set may become stale. DNTG therefore periodically refreshes the negative target set:

$$\mathcal{N}_t = \begin{cases} \Phi_K(x_t^{adv}; f_s, y), & \text{if } t \equiv 0 \pmod{R}, \\ \mathcal{N}_{t-1}, & \text{otherwise}, \end{cases} \tag{10}$$

where R is the refresh interval. This refresh mechanism allows the guidance direction to adapt to the changing adversarial state.

After obtaining $\mathcal{N}_t$, DNTG constructs a temperature-controlled soft negative target distribution over the selected classes:

$$q_t(c) = \frac{\mathbf{1}[c \in \mathcal{N}_t] \exp\left(z_s^c(x_t^{adv})/T\right)}{\sum_{j \in \mathcal{N}_t} \exp\left(z_s^j(x_t^{adv})/T\right)}, \quad c \in \mathcal{C}^-(y), \tag{11}$$

where T controls the sharpness of the negative target distribution. A smaller T emphasizes the most confident negative classes, while a larger T produces smoother guidance over multiple negative targets. Thus, temperature scaling balances hard negative guidance and averaged negative guidance.

Algorithm 1 Dynamic Negative Target Guidance (DNTG)

Require: Clean skeleton x, label y, surrogate model f_s, base attack $\mathcal{A}$, iterations M, negative set size K, temperature T, refresh interval R, guidance weight λ
Ensure: Adversarial skeleton x_M^{adv}
1: Initialize $x_0^{adv} = x$ and $\mathcal{C}^-(y) = \{c \mid c \neq y\}$
2: Initialize $\mathcal{N}_0 = \Phi_K(x_0^{adv}; f_s, y)$
3: **for** $t = 0$ to $M - 1$ **do**
4: **if** $t \bmod R == 0$ **then**
5: $\mathcal{N}_t = \Phi_K(x_t^{adv}; f_s, y)$
6: **else**
7: $\mathcal{N}_t = \mathcal{N}_{t-1}$
8: **end if**
9: Build q_t by truncated softmax over $\mathcal{N}_t$ via Eq. (11)
10: Compute $\mathcal{L}_{\text{DNTG}} = \mathbb{E}_{c\sim q_t}\left[z_s^c(x_t^{adv}) - z_s^y(x_t^{adv})\right]$
11: Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{DNTG}}$
12: Update x_{t+1}^{adv} using the original update rule of base attack $\mathcal{A}$ with objective $\mathcal{L}_{\text{total}}$
13: **end for**
14: **return** x_M^{adv}

Based on this distribution, DNTG defines the negative guidance objective as an expected logit-margin over high-confidence negative classes:

$$\mathcal{L}_{\text{DNTG}}(x_t^{adv}) = \mathbb{E}_{c\sim q_t}\left[z_s^c(x_t^{adv}) - z_s^y(x_t^{adv})\right]. \tag{12}$$

Equivalently, this objective can be written as:

$$\mathcal{L}_{\text{DNTG}}(x_t^{adv}) = \sum_{c\in\mathcal{N}_t} q_t(c)\left(z_s^c(x_t^{adv}) - z_s^y(x_t^{adv})\right). \tag{13}$$

This formulation does not force the adversarial sample toward a single fixed target class. Instead, it maximizes the expected margin between multiple informative negative classes and the ground-truth class, which provides both directional guidance and optimization stability.

DNTG is added to the original attack objective as a plug-in guidance term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda\mathcal{L}_{\text{DNTG}}, \tag{14}$$

where λ controls the strength of negative target guidance. The corresponding optimization direction can be decomposed as:

$$\nabla_{x_t^{adv}}\mathcal{L}_{\text{total}} = \nabla_{x_t^{adv}}\mathcal{L}_{\text{base}} + \lambda\nabla_{x_t^{adv}}\mathcal{L}_{\text{DNTG}}. \tag{15}$$

This decomposition shows that DNTG preserves the original optimization structure of the base attack while injecting an additional transferable negative-class guidance direction.

For iterative gradient-based attacks, the adversarial sample is updated by:

$$x_{t+1}^{adv} = \Pi_{\mathcal{B}_\epsilon(x)}\left(x_t^{adv} + \alpha\cdot\operatorname{sign}\left(\nabla_{x_t^{adv}}\mathcal{L}_{\text{total}}\right)\right), \tag{16}$$

where $\Pi_{\mathcal{B}_\epsilon(x)}(\cdot)$ denotes projection onto the feasible perturbation set and α is the step size.

The complete iterative execution flow of DNTG is formally summarized in Algorithm 1. In each iteration t, the algorithm first determines whether to refresh the high-confidence negative set $\mathcal{N}_t$ based on the specified interval R (Lines 4–8). Next, it builds the temperature-scaled soft guidance distribution $q_t(c)$ over $\mathcal{N}_t$ via Eq. (11) and computes the negative guidance loss $\mathcal{L}_{\text{DNTG}}$ (Lines 9–10). Finally, DNTG combines this with the baseline objective to form $\mathcal{L}_{\text{total}}$, updating the adversarial skeleton x_{t+1}^{adv} within the bounded perturbation set $\mathcal{B}_\epsilon(x)$ (Lines 11–12). Through negative target mining, temperature-scaled reweighting, and dynamic trajectory refresh, DNTG achieves a harmonious balance among directionality, stability, and adaptability for transferable black-box skeleton attacks.

4 Experiments

In this section, we evaluate the effectiveness of the proposed DNTG by integrating it into six iterative skeleton adversarial attack algorithms across multiple datasets and target models.

4.1 Experimental Setup

Datasets and models. We evaluate DNTG on NTU RGB+D 60 (Shahroudy et al., 2016), NTU RGB+D 120 (Liu et al., 2019), and HDM05 (Müller et al., 2007). NTU RGB+D 60 and NTU RGB+D 120 are used for the main black-box transfer evaluation, while HDM05 is used to examine cross-dataset generalization under different motion distributions. We use ST-GCN (Yan et al., 2018) and CTR-GCN (Chen et al., 2021) as surrogate models, and evaluate the generated adversarial examples on heterogeneous target models, including 2s-AGCN (Shi et al., 2019), MS-G3D (Liu et al., 2020), FR-Head (Zhou et al., 2023), and the corresponding non-surrogate model among ST-GCN and CTR-GCN. This target pool covers representative graph-based skeleton recognition architectures with different spatial-temporal modeling designs.

Attack baselines and metrics. DNTG is integrated into six representative skeleton attack methods: I-FGSM (Kurakin et al., 2018), MI-FGSM (Dong et al., 2018), SMART (Wang et al., 2021), CIASA (Liu et al., 2022), DIM (Xie et al., 2019), and TASAR (Diao et al., 2025). For each baseline, we compare the original attack with its DNTG-enhanced variant under the same perturbation budget and optimization setting. Unless otherwise specified, all attacks are optimized for 200 iterations with perturbation budget $\epsilon = 0.01$. DNTG uses $K = 10$ negative target classes, temperature $T = 10$, refresh interval $R = 10$, and guidance weight $\lambda = 3.5$ by default. All experiments are conducted on NVIDIA A100 GPUs. We report Attack Success Rate (ASR, %), where higher values indicate stronger black-box transferability. The surrogate model itself is excluded when computing average ASR.

4.2 Comparison of Attack Results

4.2.1 Results on NTU60 and NTU120

We first evaluated the overall effectiveness of DNTG on NTU RGB+D 60 and NTU RGB+D 120. Using ST-GCN (Yan et al., 2018) and CTR-GCN (Chen et al., 2021) as surrogate models, we generated adversarial examples with six baseline attacks and transferred them to multiple heterogeneous target models. The results are summarized in Table 1.

Overall, DNTG improved the average ASR for most attack methods and surrogate-model settings. This indicates that the proposed guidance mechanism is not tied to a specific attack objective or surrogate architecture. The improvement is more evident on NTU60, while NTU120 remains more challenging due to its larger number of action categories and finer-grained motion semantics. Nevertheless, DNTG still achieved consistent gains in most NTU120 settings, suggesting that high-confidence negative classes can provide useful guidance under different dataset scales and category complexities.

A closer comparison across attack methods shows that DNTG brings particularly clear improvements to SMART (Wang et al., 2021) and CIASA (Liu et al., 2022). These attacks primarily emphasize motion plausibility constraints or ground-truth confidence suppression, and therefore provide limited explicit modeling of cross-model category guidance. By introducing high-confidence negative classes, DNTG supplies an additional optimization direction beyond simply moving away from the ground-truth class, leading to improved ASR across multiple target models.

For stronger baselines such as I-FGSM (Kurakin et al., 2018), MI-FGSM (Dong et al., 2018), DIM (Xie et al., 2019), and TASAR (Diao et al., 2025), DNTG also improves the average ASR in most cases, although the gains are relatively moderate. This suggests that DNTG acts as a complementary output-level guidance mechanism rather than replacing the original attack objectives. Taken together, Table 1 demonstrates that DNTG provides a general transferability enhancement on the NTU benchmarks.

4.2.2 Results on HDM05

We further investigate whether the effectiveness of DNTG depends on the distribution characteristics of specific datasets, such as the NTU series datasets. To evaluate its generalization ability under different data distributions, we apply DNTG to the HDM05 dataset for cross-dataset evaluation. The results are summarized in Table 2.

Table 1: Black-box attack success rate (ASR, %) on NTU RGB+D 60 and NTU RGB+D 120. DNTG is integrated into different baseline attacks. Higher values indicate better transferability. The better result within each baseline–DNTG pair is highlighted in bold.

		ST-GCN as Proxy Model					CTR-GCN as Proxy Model				
Dataset	**Method**	**AGCN**	**MSG3D**	**CTR-GCN**	**FR-Head**	**Avg.**	**ST-GCN**	**AGCN**	**MSG3D**	**FR-Head**	**Avg.**
NTU 60	I-FGSM	43.59	72.79	16.24	40.03	43.16	34.38	57.81	72.36	59.98	56.13
	+DNTG	**47.15**	**77.21**	**17.09**	**42.88**	**46.08**	**34.98**	**58.89**	**79.69**	**64.90**	**59.62**
	MI-FGSM	45.16	70.66	15.81	41.31	43.23	34.62	57.69	70.31	58.77	55.35
	+DNTG	**48.72**	**74.79**	**16.38**	**41.74**	**45.41**	**36.66**	**59.13**	**77.52**	**64.06**	**59.34**
	SMART	6.98	3.85	1.99	4.84	4.42	17.79	15.38	7.81	10.94	12.98
	+DNTG	**10.54**	**6.41**	**3.42**	**7.12**	**6.87**	**20.79**	**20.43**	**12.26**	**20.07**	**18.39**
	CIASA	6.98	3.85	1.99	4.70	4.38	17.43	15.02	7.81	10.58	12.71
	+DNTG	**10.83**	**6.55**	**3.56**	**7.26**	**7.05**	**20.79**	**20.55**	**12.26**	**20.55**	**18.54**
	DIM	31.20	33.62	**15.38**	25.93	26.53	**32.57**	57.09	66.83	61.06	54.39
	+DNTG	**34.05**	**38.46**	14.25	**27.35**	**28.53**	**32.57**	**58.77**	**67.07**	**61.30**	**54.93**
	TASAR	46.51	**74.88**	14.88	39.53	43.95	30.80	54.00	66.40	58.00	52.30
	+DNTG	**49.30**	74.42	**17.67**	**40.93**	**45.58**	**32.40**	**57.20**	**67.20**	**59.60**	**54.10**
NTU 120	I-FGSM	**23.44**	26.72	14.37	31.09	23.91	22.13	**32.78**	40.77	**28.45**	31.03
	+DNTG	23.28	**29.53**	**14.69**	**32.34**	**24.96**	**23.46**	31.45	**44.59**	**28.45**	**31.99**
	MI-FGSM	22.34	26.41	15.62	30.63	23.75	20.13	30.95	41.26	28.45	30.20
	+DNTG	**22.97**	**29.06**	**16.56**	**31.87**	**25.12**	**25.12**	**33.94**	**45.09**	**30.28**	**33.61**
	SMART	12.03	10.00	6.88	9.69	9.65	9.65	11.65	8.82	9.15	9.82
	+DNTG	**15.31**	**12.34**	**9.22**	**11.72**	**12.15**	**12.48**	**13.81**	**11.81**	**11.65**	**12.44**
	CIASA	12.03	10.47	7.03	9.53	9.77	7.65	11.48	8.82	9.15	9.28
	+DNTG	**15.62**	**12.81**	**9.84**	**11.72**	**12.50**	**12.48**	**13.98**	**12.15**	**11.81**	**12.61**
	DIM	22.50	27.50	13.91	24.22	22.03	17.97	30.78	38.27	22.63	27.41
	+DNTG	**29.06**	**30.94**	**20.94**	**32.19**	**28.28**	**22.96**	**38.94**	**44.43**	**34.28**	**35.15**
	TASAR	21.20	32.00	15.60	**31.20**	25.00	16.60	31.06	**42.55**	28.94	29.79
	+DNTG	**22.40**	**35.60**	**16.80**	30.80	**26.40**	**18.30**	**34.04**	**42.55**	**29.79**	**31.17**

Table 2: Black-box transfer attack results on HDM05. The **Avg.** column reports the mean Attack Success Rate (ASR, %) over all victim models. The better result within each baseline–DNTG pair is highlighted in bold.

		ST-GCN as Proxy Model					CTR-GCN as Proxy Model				
Dataset	**Method**	**CTR-GCN**	**FR-Head**	**MSG3D**	**AGCN**	**Avg.**	**ST-GCN**	**FR-Head**	**MSG3D**	**AGCN**	**Avg.**
HDM05	I-FGSM	38.46	14.67	27.41	**34.52**	28.77	35.34	12.16	8.33	8.54	16.09
	+DNTG	**41.79**	**14.88**	**27.82**	33.05	**29.39**	**40.75**	**14.64**	**10.21**	**11.25**	**19.21**
	MI-FGSM	39.09	13.64	**28.03**	**35.15**	28.98	33.26	12.78	8.54	9.17	15.94
	+DNTG	**43.04**	**15.70**	**28.03**	33.89	**30.16**	**39.50**	**13.61**	**8.96**	**11.67**	**18.43**
	SMART	0.42	0.00	0.63	0.42	0.37	3.95	1.24	0.62	0.21	1.51
	+DNTG	**3.33**	**3.51**	**2.30**	**2.72**	**2.97**	**13.72**	**6.60**	**1.25**	**2.92**	**6.12**
	CIASA	0.42	0.00	0.63	0.42	0.37	3.74	1.24	0.62	0.21	1.45
	+DNTG	**3.33**	**3.51**	**2.30**	**2.51**	**2.91**	**12.06**	**6.60**	**1.46**	**2.71**	**5.71**
	DIM	37.84	13.22	19.25	42.89	28.30	23.49	**17.11**	4.38	**11.46**	14.11
	+DNTG	**38.67**	**13.84**	**19.87**	**43.10**	**28.87**	**24.53**	16.49	**5.42**	10.42	**14.21**
	TASAR	33.89	10.74	25.52	**35.56**	26.43	35.34	12.99	8.96	12.92	17.55
	+DNTG	**41.16**	**16.12**	**29.08**	34.52	**30.22**	**38.67**	**14.02**	**9.17**	**13.12**	**18.75**

On HDM05, DNTG still improves the average ASR for most attack baselines under both ST-GCN and CTR-GCN surrogates. This result indicates that high-confidence negative classes do not merely reflect incidental prediction biases on a specific dataset, but can provide category-level transfer signals under different motion distributions.

4.3 Ablation Study

To analyze the source of DNTG's improvement, we conduct an ablation study on negative target guidance and dynamic refresh. The results are reported in Table 3. Compared with the Base setting that only uses the original attack objective, introducing high-confidence negative classes as guidance signals improves the mean ASR from 51.15% to 52.04%. This indicates that negative target guidance provides additional category-level optimization directions beyond ground-truth suppression. When dynamic refresh is further incorporated, the full DNTG achieves the highest mean ASR of 53.60%.

Table 3: Ablation of dynamic negative target guidance. "Neg. Target" applies high-confidence non-target classes for guidance; "Refresh" periodically updates the negative target set.

Method	Neg. Target	Refresh	Mean ASR (%) ↑
Base	×	×	51.15
w/o Refresh	✓	×	52.04
Full DNTG	✓	✓	**53.60**

Table 4: Dynamic evolution of negative target sets. **Jac.-Init** and **Jac.-Prev** denote the Jaccard similarity of the current top-K set with the initial and previous sets, respectively. **Top-1 Sw.** denotes the ratio of samples whose dominant negative target changes from initialization.

Iteration	Jac.-Init	Jac.-Prev	Top-1 Sw.
0	1.0000	1.0000	0.0000
10	0.7677	0.7677	0.2837
50	0.5367	0.9460	0.4808
100	0.4902	0.9821	0.5120
150	0.4708	0.9917	0.5337
200	0.4611	0.9956	0.5433

These results show that the benefit of DNTG comes not only from using a static negative target set, but also from dynamically updating this set during optimization. As adversarial samples continuously move away from their initial states, the surrounding prediction distribution and negative-class decision structure may also change. If the initial negative target set is kept fixed throughout optimization, the attack may be constrained by outdated guidance directions. Dynamic refresh updates the high-confidence negative classes according to the current adversarial state, allowing the guidance direction to better adapt to the evolving attack trajectory.

4.4 Dynamic Negative Target Evolution

Beyond the component-level ablation, Table 4 examines whether the selected negative target set remains fixed during optimization. Jac.-Init decreases from 1.0000 to 0.4611, while Top-1 Sw. increases from 0.0000 to 0.5433, indicating that the negative target set gradually departs from its initial state. Meanwhile, Jac.-Prev increases to 0.9956 in later iterations, showing that the target set becomes stable rather than oscillating randomly. These results support the use of dynamic refresh, which adapts the guidance targets to the evolving adversarial state while maintaining stable guidance in later stages.

Additional analyses on hyperparameter sensitivity, qualitative visualization, and computational overhead are provided in Appendix A.

5 Conclusion

In this paper, we revisit untargeted transfer attacks on skeleton-based action recognition and highlight the importance of explicitly exploiting non-ground-truth responses for transfer-oriented optimization. Motivated by the observed cross-model consistency of high-confidence negative classes, we propose Dynamic Negative Target Guidance (DNTG), a plug-and-play framework that dynamically mines and reweights informative negative targets throughout attack optimization. This design provides adaptive negative-class guidance while reducing dependence on surrogate-specific responses. Extensive experiments across three benchmarks demonstrate that DNTG improves average black-box transferability across multiple attack baselines and representative skeleton recognition models.

REFERENCES

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017. doi: 10.1109/SP.2017.49.

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13359–13368, 2021.

Yunfeng Diao, Tianjia Shao, Yong-Liang Yang, Kun Zhou, and He Wang. Basar: Black-box attack on skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7597–7607, 2021.

Yunfeng Diao, He Wang, Tianjia Shao, Yongliang Yang, Kun Zhou, David Hogg, and Meng Wang. Understanding the vulnerability of skeleton-based human activity recognition via black-box attack. *Pattern Recognition*, 153:110564, 2024.

Yunfeng Diao, Baiqi Wu, Ruixuan Zhang, Ajian Liu, Xiaoshuai Hao, Xingxing Wei, Meng Wang, and He Wang. TASAR: Transfer-based attack on skeletal action recognition. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition. In *Computer Vision – ECCV 2024*, pp. 401–420, Cham, 2025. Springer Nature Switzerland.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Zi Kang, Hui Xia, Rui Zhang, Shuliang Jiang, Xiaolong Shi, and Zuming Zhang. Fgda-gs: Fast guided decision attack based on gradient signs for skeletal action recognition. *Computers & Security*, 135:103522, 2023.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Jian Liu, Naveed Akhtar, and Ajmal Mian. Adversarial attack on skeleton-based human action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1609–1622, 2022.

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2019.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 143–152, 2020.

Zhengzhi Lu, He Wang, Ziyi Chang, Guoan Yang, and Hubert P. H. Shum. Hard no-box adversarial attack on skeleton-based human action recognition with skeleton-motion-informed gradient. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4597–4606, 2023.

Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, 2007.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. *arXiv preprint arXiv:1602.02697*, 2016b.

Helei Qiu, Biao Hou, Bo Ren, and Xiaohua Zhang. Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv preprint arXiv:2201.02849*, 2022.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1010–1019, 2016.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12026–12035, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Nariki Tanaka, Hiroshi Kera, and Kazuhiko Kawamoto. Adversarial bone length attack on action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2335–2343, 2022.

He Wang, Feixiang He, Zhexi Peng, Tianjia Shao, Yong-Liang Yang, Kun Zhou, and David Hogg. Understanding the robustness of skeleton-based action recognition under adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14656–14665, 2021.

Juanjuan Weng, Zhiming Luo, Shaozi Li, Nicu Sebe, and Zhun Zhong. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration. *IEEE Transactions on Information Forensics and Security*, 18:3561–3574, 2023. doi: 10.1109/TIFS.2023.3284649.

Zhize Wu, Pengpeng Sun, Xin Chen, Keke Tang, Tong Xu, Le Zou, Xiaofeng Wang, Ming Tan, Fan Cheng, and Thomas Weise. Selfgcn: Graph convolution network with self-attention for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 33:4391–4403, 2024.

Zhize Wu, Yue Ding, Long Wan, Teng Li, and Fudong Nian. Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition. *Pattern Recognition*, 159: 111106, 2025.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, and In So Kweon. Investigating top-k white-box and transferable black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15085–15094, 2022.

Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10608–10617, 2023.

AI USE STATEMENT

Generative AI tools assisted with translation from Chinese into English, language editing, and organization of the method and experiment sections. The authors checked the equations, results, citations, and technical claims against the source manuscript and take responsibility for the final paper.

A ADDITIONAL EXPERIMENTAL ANALYSIS

This appendix provides additional experimental analyses of DNTG, including hyperparameter sensitivity, qualitative visualization of adversarial skeleton sequences, and computational overhead. These results complement the main experiments by further examining the robustness of DNTG to different hyperparameter choices, the visual quality of generated adversarial examples, and the additional computational cost introduced by the proposed guidance.

A.1 HYPERPARAMETER SENSITIVITY

We investigate the sensitivity of DNTG to three key hyperparameters: the guidance weight λ, the number of negative targets K, and the temperature T. The results are shown in Fig. 4.

For the guidance weight λ, the average attack success rate (ASR) increases substantially as λ grows from a small value and becomes relatively stable when λ approaches 3. This indicates that a moderate negative guidance strength is sufficient to provide most of the transferability gain, while excessively increasing its contribution brings limited additional improvement.

For the number of negative targets K, the performance remains relatively stable over a broad range of values. This suggests that DNTG does not depend strongly on a particular choice of K and can effectively exploit multiple high-confidence negative classes to construct transferable guidance.

For the temperature T, moderate values generally achieve better performance. When T is too small, the negative target distribution becomes overly concentrated on a few dominant classes, which may increase the risk of overfitting to surrogate-specific responses. In contrast, an excessively large T produces an overly smooth distribution and weakens the discriminative differences among negative targets. Based on these observations, we use $\lambda = 3.5$, $K = 10$, and $T = 10$ as the default configuration in our experiments.

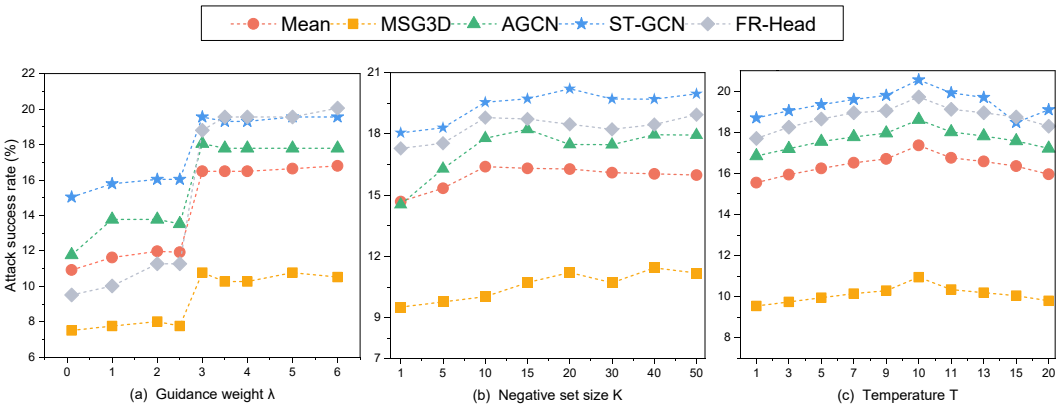


Figure 4: Sensitivity analysis of DNTG with respect to three key hyperparameters: (a) the guidance weight λ; (b) the number of negative targets K; and (c) the temperature T.

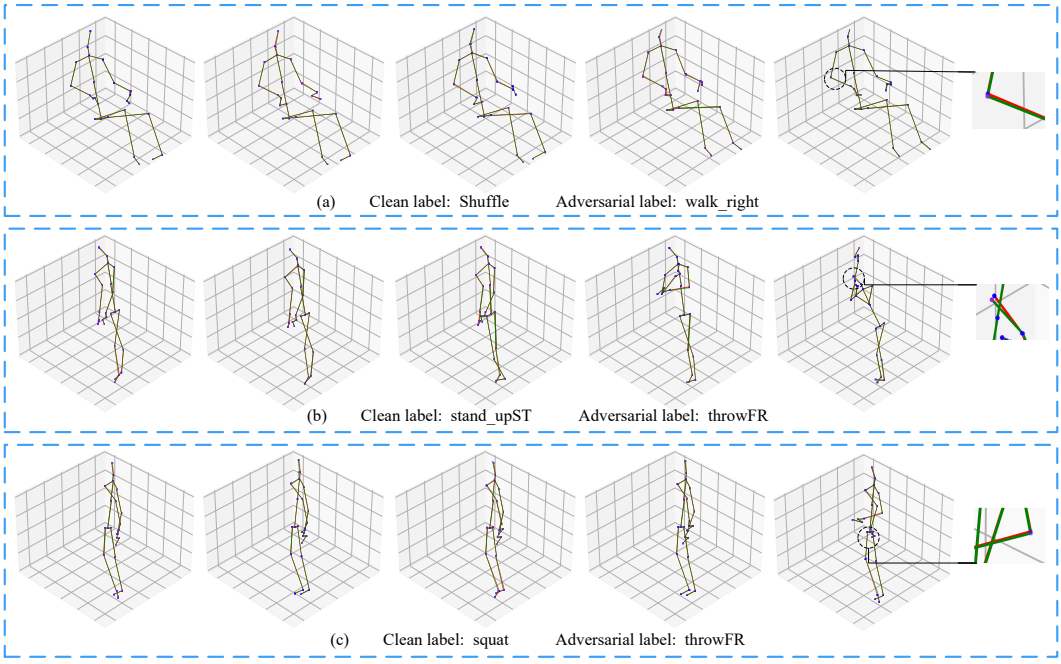


Figure 5: Qualitative visualization of adversarial skeleton sequences generated by SMART+DNTG for three representative action categories. Green and red skeletons denote clean and adversarial samples, respectively.

A.2 Qualitative Analysis

We further provide qualitative examples to examine whether the improved transferability of DNTG is achieved at the cost of severe skeleton distortion. Figure 5 visualizes clean and adversarial skeleton sequences generated by SMART+DNTG for three representative action categories.

As shown in Fig. 5, the adversarial skeleton sequences largely preserve the global body structure and temporal motion patterns of the corresponding clean samples. Although the generated perturbations successfully alter the model predictions, the differences between clean and adversarial skeletons remain visually subtle. These results indicate that the transferability improvement introduced by DNTG does not rely on severe deformation of the original skeletal motion.

A.3 Computational Cost

Table 5: Runtime and GPU memory overhead of DNTG.

Attack	**Time (s)**		**Memory (MB)**	
	Base	**+DNTG**	**Base**	**+DNTG**
MI-FGSM	7	13	8093.28	8094.74
SMART	28	44	8103.74	8177.86

We additionally evaluate the computational overhead introduced by DNTG in terms of runtime and GPU memory consumption. Since DNTG operates as a plug-and-play guidance mechanism and does not modify the architecture or parameters of the surrogate model, its additional memory requirement is limited.

As reported in Table 5, incorporating DNTG into MI-FGSM increases the runtime from 7 s to 13 s, while the GPU memory consumption changes only slightly from 8093.28 MB to 8094.74 MB. For SMART, the runtime increases from 28 s to 44 s, while the GPU memory consumption increases

from 8103.74 MB to 8177.86 MB. The additional runtime mainly results from the computation and periodic update of the dynamic negative target guidance during iterative optimization. Overall, DNTG introduces limited additional memory overhead while remaining compatible with existing attack pipelines.